Bayesian Inference and Model Selection for Partially Observed Stochastic Epidemics

by

Panayiota Touloupou

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

October 2016

THE UNIVERSITY OF WARWICK

Στην οικογένειά μου.

“ Να λες πάντα αυτό που νιώθεις και
να κάνεις πάντα αυτό που σκέφτεσαι.”

“ Κανείς δεν θα σε θυμάται για τις
κρυφές σου σκέψεις.”

– **Γκαμπριέλ Γκαρσία *Μάρκες***

Contents

List of Tables

LIST OF FIGURES

Acknowledgments

First of all I would like to thank my supervisor, Dr Simon Spencer, for his outstanding guidance, motivation and enthusiasm over the past four years. I consider myself lucky to have been given the change to work with him. Also, I am grateful to my second supervisor, Prof Bärbel Finkenstädt, for the invaluable advice and feedback regarding my work. Only because of them, I have managed to make it this far.

Special thanks goes to the members of the Greek-Cypriot Earlsdon community consisting of Pantelis, Thodoris, Giorgos, Kyriaki and little Lukia, for their tremendous support and for making my free time much more enjoyable. I am grateful to all my PhD colleagues and friends and in particular to Silvia, Simone, Giacomo, Habib, Fiona, Ioanna, Helen and Murray for all the great times that we had together and for the extraordinary help that they have provided. Moreover, I wish to thank all the people in the Warwick Statistics department for being so kind and for creating such a stimulating academic environment from which I have utterly benefited.

To my family for putting so much faith in me and for being by my side in all ups and downs. Last but not least, I feel deeply indebted to my close friends Eirini, Sofia, Agathi and Andrea from Cyprus who have been constantly encouraging me; I feel sorry for missing so many important moments and I promise to make up for it.

Finally, I am grateful to the Warwick Statistics department for funding my PhD studies.

Declarations

I hereby declare that this thesis is the result of my own work, except where stated otherwise. The thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy and has not been submitted in any other application for any degree.

- Chapter 4 is collaborative work with Naif Alzahrani, Peter Neal, Simon E.F. Spencer and Trevelyan J. McKinley. The material presented in this chapter is based on the following manuscript which was recently submitted for publication:
  Touloupou, P., Alzahrani, N., Neal, P., Spencer, S. E., and McKinley, T. J. (2015). Model comparison with missing data using MCMC and importance sampling. *arXiv preprint arXiv:1512.04743.*

- The works presented in Chapters 5, 6 and 7 will be submitted for publication soon.

Panayiota Touloupou, September 2016

ABSTRACT

Over the past decades, statistical models have been established as an important tool for understanding the transmission dynamics of infectious diseases. Inference in such models can be challenging due to the strong dependencies in the actual epidemic process, as well as the fact that observations often rely in diagnostic tests that have imperfect sensitivities. Moreover, samples are often taken with very low temporal resolution, which leads to the actual dynamics being only partially observed. Data augmentation techniques implemented within the framework of Markov chain Monte Carlo (MCMC) methods can tackle these problems by taking into account the unobserved dynamics of transmission and thus have been widely employed in practice. Despite the methodological advances in the context of partially observed epidemic models, there are still several open challenges that remain to be addressed. One of the key challenges is the establishment of model comparison techniques that can be efficiently applied in problems involving a large amount of missing information. In this thesis, we describe a framework based on importance sampling which provides estimates of the marginal likelihood and is well suited for applications in this complex setting. Until recently, the study of infectious diseases in large scale populations has been challenging due to the computationally intensive methods needed to these models. One further contribution of this thesis is the development of a data augmentation MCMC algorithm that can be used in both Markovian and non-Markovian epidemic models. Our algorithm achieves good computational efficiency and therefore can be viewed as an alternative to existing approaches, particularly for applications on big datasets. The last part of the thesis is concerned with epidemic data containing additional information regarding the strain of a pathogen with which individuals are infected. Quantifying the interactions between the different strains of pathogens is crucial in order to obtain a complete understanding of the disease but statistical methods for this type of problem are still in the early stages of development. Motivated by this demand, we construct a model that incorporates this additional information and propose a statistical algorithm for inference. The model improves upon existing methods in the sense that it allows for both imperfect diagnostic test sensitivities and strain misclassification. Finally, extensive simulation studies are conducted in order to assess the performance of our methods, while the utility of the developed methodologies is demonstrated on data obtained from two longitudinal studies of *Escherichia coli* in cattle.

Abbreviations

ABC	Approximate Bayesian Computation
ACF	Autocorrelation Function
AUC	Area Under Curve
BS	Bridge Sampling
CHMM	Coupled Hidden Markov Model
CHSMM	Coupled Hidden Semi-Markov Model
CI	Credible Interval
E. coli	*Escherichia Coli*
EM	Expectation-Maximisation
ESS	Effective Sample Size
FFBS	Forward Filtering Backwards Sampling
HIV	Human Immunodeficiency Virus
HM	Harmonic Mean
HMC	Hamiltonian Monte Carlo
HMM	Hidden Markov Model
iid	Independent and Identically Distributed
iFFBS	Individual Forward Filtering Backwards Sampling
IS	Importance Sampling
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings

PCR	Polymerase Chain Reaction
PFGE	Pulsed-Field Gel Electrophoresis
Pnc	Pneumococcal Nasopharyngeal Carriage
PP	Power Posterior
RAJ	Recto-Anal Junction
RAMS	Recto-Anal Mucosal Swab
RJcor	Reversible Jump Markov Chain Monte Carlo Corrected
RJMCMC	Reversible Jump Markov Chain Monte Carlo
ROC	Receiver Operating Characteristics
S.D.	Standard Deviation
SEIR	Susceptible-Exposed-Infected-Removed
SIR	Susceptible-Infected-Removed
SIS	Susceptible-Infected-Susceptible
SMC	Sequential Monte Carlo
tESS	Time Normalised Effective Sample Size

Chapter 1

Introduction

1.1 Epidemic background

1.1.1 The need for epidemic modelling

Infectious diseases have been historically, and remain, one of the major causes of human mortality and suffering. For instance, the Black Death pandemic of the 14th century killed about one fourth of the population in Europe, while in 1520 half of the population of Aztecs lost their lives due to smallpox (Bailey, 1975). Another notable example is the global spread of the human immunodeficiency virus (HIV), which nearly 35 years after its first reported case continues to be one of the main causes of death worldwide. In 2014, 1.2 million died of illnesses related to HIV and it is believed that there were 2 million new infections (UNAIDS, 2015). More recently, the influenza pandemic caused by the H1N1 virus spread rapidly to over 30 countries during the first few weeks of its surveillance (Smith *et al.*, 2009), leading to over 18,000 deaths[1]. At present, the World Health Organisation estimates in its latest report that over 6 million people die annually due to infectious and parasitic diseases, which represents 10% of the total deaths per year[2].

The examples above highlight the need for epidemic models in order to gain a better understanding of the transmission dynamics of an infectious disease. Mathematical models allow us to capture the features that drive the spread of the disease and obtain estimates of several important epidemiological quantities. These can be subsequently used to predict the progression of an epidemic and hence guide the development of real-time control strategies to prevent a potential outbreak (see e.g. Bailey, 1975; Keeling and Rohani, 2008). Given a particular question of interest, epidemic models can further contribute to formulate what data should be collected in order to answer this question (Isham, 2005).

[1] Pandemic (H1N1) 2009 - update 103. Disease Outbreak News. World Health Organization. 2010-06-04. (`http://www.who.int/csr/don/2010_06_04/en/`)

[2] Global Health Estimates 2013: deaths by cause, age and sex; estimates for 2000–2012. Geneva: World Health Organization; 2014. (`http://www.who.int/healthinfo/global_burden_disease/en/`)

1.1.2 History of epidemic modelling

The history of mathematical models for infectious diseases starts in 1760 with a paper by Bernoulli (1760), who developed a model to evaluate the effectiveness of inoculation against the smallpox virus. However, it was not until the 20th century that the field was established. In 1906 Hamer (1906) studied a discrete time deterministic model for measles epidemics, introducing the principle of *mass action* or *homogeneous mixing* which assumes that the probability of a new infection in the next time point is proportional to the product of the total number of susceptible and infected individuals in the population. This idea was then extended to the continuous time case with the works by Ross (1911, 1916), Ross and Hudson (1917a,b) and Kermack and McKendrick (1927). In the latter, Kermack and McKendrick proposed the first complete mathematical model, which is known as the deterministic *general epidemic model.*

During the same period, some stochastic models were developed in order to account for the randomness observed in real-life epidemics, starting with the continuous time model of McKendrick (1926) which is a stochastic variant of the general epidemic model. Another model that received considerable attention was the discrete time chain-binomial model proposed by Reed and Frost for the purposes of a series of lectures given in 1928, see e.g. Abbey (1952). In the Reed-Frost model the number of infected individuals at any given time point has a binomial distribution with the probability of infection depending on the number of carriers at the previous time point. In a landmark paper, Bartlett (1949) studied McKendrick's model and since then the literature on stochastic epidemic models began to grow exponentially. Some textbook references include Bailey (1975), Becker (1989) and Andersson and Britton (2000), whereas Isham (2005) and Britton (2010) provide comprehensive reviews of the topic.

Several studies within the aforementioned references classify individuals into compartments according to their disease states. Such approaches are referred to as *compartmental* models. The most characteristic example within this class is the Susceptible-Infected-Removed model (SIR, McKendrick, 1926; Whittle, 1955; Barbour, 1975, for example) where individuals can be *susceptible* if they do not have the disease but can acquire it, *infected* if they carry the disease and can transmit it to susceptible individuals or *removed* when they have recovered from the disease and cannot be infected by it again. There exist many extensions of the SIR model, one of the most widely used being the SEIR model (see e.g. Bartlett, 1956; Gibson and Renshaw, 1998; Lekone and Finkenstädt, 2006) in which the additional *exposed* state represents individuals who have the disease but are not infectious. Another example is the SIS model (Bartlett, 1957; Weiss and Dishon, 1971; Nåsell, 2002, among others) which assumes that cleared individuals become available for re-infection immediately after recovery.

The majority of the works listed so far assumes a community of homogeneous individuals mixing uniformly. Nevertheless, this is often unrealistic in practice due to variations induced by the characteristics of the population or the disease. As a result, significant efforts have been made towards relaxing this assumption by accounting for different sources of heterogeneity. Individual heterogeneities allow for the members of the population to have different risks of becoming infected or transmitting the disease according to their attributes. For instance, children are more susceptible to influenza compared to adults. Numerous other factors exist, including gender (especially for sexually transmitted diseases), vaccination status and previous exposure to the disease. These factors can be used to categorise individuals into different types, assuming that individuals of the same type exhibit identical behaviour (Becker and Marschner, 1990; Andersson and Britton, 1998; Hayakawa *et al.*, 2003). Similar ideas can be applied to model heterogeneities in the parasite itself, which occur for example when several strains of the same virus are observed (Ball and Clancy, 1995; Ferguson *et al.*, 2003). In such cases it is essential to model interactions between strains since infection by one type may lead to immunity or partial immunity to the other types of the disease.

Another source of heterogeneity may arise when the population under study is organised into small groups, for example households or schools. Typically the mixing within these groups occurs at a higher rate than with the rest of the population, and therefore it is of vital importance to take this structure into consideration. In the majority of the applications it is plausible to assume that individuals mix homogeneously within each group. The simplest approach is to exclude the possibility of infection from the community and focus on the dynamics of the disease within each household. An alternative approach is to assume that each individual avoids infection from the population at large (global infection) with a single fixed probability, implying that households are independent of one another (Longini and Koopman, 1982). In some scenarios it is more realistic to allow for two levels of mixing, where infectious contacts take place both at a local (within group) level and at a global level (Ball *et al.*, 1997; Demiris and O'Neill, 2005). The difference with the previous model is that households are no longer independent and in particular the probability of avoiding global infection depends on the number of infected individuals in the community. A generalisation of the model with two levels of mixing is the multilevel model where individuals belong to more than one group simultaneously, see for example Cauchemez *et al.* (2008) and Britton *et al.* (2011).

Another direction that researchers have taken in order to make models more accurately represent the realities of disease progression concerns the infection period that is, the time that an individual remains infected. The general stochastic epidemic model uses the Markov property which states that the probability of recovery is constant over time. This is achieved by assuming that the infection period follows an Exponential or a Geometric distribution in continuous and discrete time, respectively. However, in some diseases individuals' chances of being cleared increase with the time of infection, for example because they develop immunity. Considerable progress has been made in relaxing the Markovian assumption, see for example O'Neill and Becker (2001) who used a Gamma distribution for the infection period and Streftaris and Gibson (2004) who adopted the Weibull distribution.

There are various other ways in which epidemic models can be extended. For instance, one can use geographical information regarding the incidences of a disease known as spatial epidemiology; we refer the reader to Lawson (2013) and Lawson *et al.* (2016) for an extensive treatment of the topic. A number of approaches in the literature utilise network theory to represent the interactions within a population during an epidemic; see for example Anderson *et al.* (1999), Danon *et al.* (2011) and the references therein. However, such models are not further discussed here.

1.1.3 Statistical inference and model selection for stochastic epidemic models

So far, we have focused on reviewing some of the existing work on modelling the dynamics of an infectious disease. From this point on, we restrict our attention on stochastic epidemic models which are the key theme of the thesis. In this section, we summarise available tools for drawing inference about parameters in such models. Inference in epidemics is not a trivial problem and therefore requires a special methodology. One of the main difficulties is the existence of dependencies in the data that arise because of the contacts made between individuals. A further complication is the fact that the actual process of infection is in most cases only partially observed, in the sense that times of acquiring and clearing infection are not known exactly. Moreover, the tests that are used to detect the disease may be imperfect, thus leading to data of lower quality. For all these reasons, it is often difficult to analytically evaluate the likelihood because its calculation involves integrating out all unobserved quantities.

When the full data are available, i.e. the times of infection and recovery are known, one can use standard techniques to obtain estimates of the parameters of interest. One example is given by maximum likelihood methods, see e.g. Becker (1989). However, since data are typically incomplete, most of the literature deals with methods that tackle the problem of inference in partially observed epidemics. Initial approaches use martingale theory to obtain method of moments estimates for the model parameters (Becker, 1989; Rida, 1991; Becker and Hasofer, 1997). Nevertheless, it is hard to extend these methods to the complex models that are used in practice. Instead, it is more common to employ data augmentation methods, which treat the missing data as additional model parameters. Such models can be handled with the Expectation-Maximisation (EM) algorithm (Becker, 1997; Becker and Britton, 1999), but are more often fitted under the Bayesian paradigm using Markov chain Monte Carlo (MCMC) methods. More details on the basic concepts of Bayesian inference and MCMC are given in Section 1.2.

The first data augmentation MCMC algorithms were developed by Gibson and Renshaw (1998) and O'Neill and Roberts (1999) for statistical analysis of the continuous time SEIR and SIR models, respectively. After that, several works adapting MCMC techniques with data imputation appeared in the literature of which we quote a few key references. O'Neill (2002) studied a simple household model with independent groups, while Demiris and O'Neill (2005) first presented a Bayesian methodology for the model with two levels of mixing. Hayakawa *et al.* (2003) extended the basic model to incorporate host heterogeneity and develop an MCMC algorithm for parameter estimation. O'Neill and Becker (2001) and Streftaris and Gibson (2004) were among the first to apply MCMC in models with a non-Markovian infection period. Smith and Vounatsou (2003) demonstrate the use of discrete time hidden Markov models (discussed in more detail in Section 1.3) for modelling longitudinal epidemiological data, which can effectively account for imperfect diagnostic tests and are used in the present work. The framework extends to partially observed continuous time epidemic models, see for example Fearnhead and Meligkotsidou (2004). Numerous other papers performed inference in epidemics with partial observations using data augmentation MCMC including Auranen *et al.* (2000), Morton and Finkenstädt (2005), Jewell *et al.* (2009), Kypraios *et al.* (2010), Erästö *et al.* (2012) and Spencer *et al.* (2015).

A class of techniques that have growing popularity in several scientific fields, along with epidemiology, are the so-called simulation-based methods. These include approximate Bayesian computation (ABC, McKinley *et al.*, 2009), sequential Monte Carlo (SMC, Ionides *et al.*, 2006; Dukic *et al.*, 2012), SMC ABC (Toni *et al.*, 2009; Toni and Stumpf, 2010) and pseudo-marginal methods (McKinley *et al.*, 2014). For a description of ABC, SMC and pseudo-marginal approaches we refer the reader to the previously mentioned papers and the references therein.

Finally, there has been some work on Bayesian model selection (see Section 1.2.4 for a brief overview) in the context of stochastic epidemic models. Initial work was based on reversible jump MCMC (RJMCMC), such as Neal and Roberts (2004) who compared alternative models for measles epidemics and O'Neill and Marks (2005) who presented an application on a *gastroenteritis* outbreak. Clancy and O'Neill (2007) demonstrated the usefulness of rejection sampling as an alternative to RJMCMC. More recently, Knock and O'Neill (2014) and O'Neill and Kypraios (2014) used path-sampling and mixture modelling, respectively, to tackle the problem. Lastly, there are examples of studies exploring the use of ABC for model choice in epidemic applications (Toni *et al.*, 2009; Lee *et al.*, 2015; Sun *et al.*, 2015).

1.1.4 Open problems

Despite the methodological advances that we described earlier in Section 1.1.3, there are still several open challenges in the area of stochastic epidemic models with partially observed data that remain to be addressed. For example, even though model selection has received fair attention, it is challenging to apply some of these methods when there are large amounts of missing data. Moreover, it is often necessary (e.g. for RJMCMC) to re-run the analysis when additional candidate models are considered leading to a significant increase in computational effort. Another issue is the lack of studies that assess the performance of existing approaches under different setups.

Even though enormous progress has been made in Bayesian data imputation techniques, most of the applications so far have been on moderate sized populations. Nevertheless, statical inference in high-dimensional missing data problems remains challenging. For example, such problems arise in individual-based models and in particular when data are gathered longitudinally from the same group of individuals for a long time period. Some noticeable developments have been made by authors such as Jewell *et al.* (2009) and Kypraios *et al.* (2010) but there is still room for improvement.

For some pathogens there exist testing procedures that can be used to distinguish among different strains. Due to the growing availability of such data, multi-strain models have increased in popularity over the past years. Inference for this class of models requires special attention due to the need to estimate several strain-specific parameters, for example acquisition or recovery rates, and to account for interactions between different strains. The fact that a carrier can be misclassified not only as being susceptible but also as being colonised by some other type rather than the true, further complicates the analysis. Although this involved setup can be handled using Bayesian techniques, there are only few tools available for practitioners.

The objective of this thesis is to attempt addressing some of the shortcomings just discussed. The methods that we propose rely upon tools from Bayesian statistics and hidden Markov models, and therefore we provide some fundamental theory behind these concepts in Sections 1.2 and 1.3, respectively. Section 1.4 contains an outline of the thesis.

1.2 Fundamentals of Bayesian methods

1.2.1 Bayes' theorem

Bayesian statistics build upon Bayes' theorem which for data $\mathbf{y}$ and model parameters $\boldsymbol{\theta}$ states:

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\pi(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\displaystyle\int_{\boldsymbol{\theta}} \pi(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}. \tag{1.1}$$

In Equation (1.1), $\pi(\boldsymbol{\theta})$ is the distribution that reflects our prior beliefs regarding the parameter vector $\boldsymbol{\theta}$, possibly high dimensional, $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ is the likelihood that follows from the model that we assume and the denominator is essentially the normalising constant. The quantity of interest is the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ that describes the updated information regarding the model parameters in light of the observed data. When available in closed form, the posterior distribution can be used to perform inference on $\boldsymbol{\theta}$. For example, the expectation of a function $f(\cdot)$ of the parameters can be calculated as:

$$\mathbb{E}\left[f(\boldsymbol{\theta})\right] = \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \mathbf{y})\,\mathrm{d}\boldsymbol{\theta}.$$

Nevertheless, posterior distributions cannot be solved analytically unless in very simple models and therefore inference is typically not straightforward. The main reason is that the integral involved in the calculations is typically intractable. One solution to this problem can be the use of Markov chain Monte Carlo methods which we briefly explain in the following Section 1.2.2.

1.2.2 Markov chain Monte Carlo

Markov chain Monte Carlo is a general technique that is used to generate samples from a distribution π, the target distribution, which is known up to a proportionality constant. The idea is to construct a Markov chain that has π as its stationary distribution and then use the chain to estimate functions of the target distribution. Within the Bayesian framework the target is the posterior distribution of the model parameters $\pi(\boldsymbol{\theta} \mid \mathbf{y})$. MCMC encompasses a broad range of algorithms; we present the ones that are relevant to our work. For a more detailed review of the topic, as well as theoretical results we refer the reader to Gilks *et al.* (1995), Roberts and Tweedie (2005) and Brooks *et al.* (2011).

1.2.2.1 The Gibbs sampler

The *Gibbs sampler* (Geman and Geman, 1984) tackles the problem of simulating from a high dimensional distribution by breaking it into a collection of lower dimensional, more manageable simulations. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$, where d is the dimension of the posterior distribution. The algorithm successively and repeatedly simulates the components θ_i from the conditional distributions $\pi(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d, \mathbf{y})$, which we call the *full conditionals.* An overview is given in Algorithm 1.

Algorithm 1: The Gibbs sampler

```
Initialise $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)}\right)$;
for $j$ = 1, 2, ..., $J$ do
    Draw $\theta_1^{(j)} \sim \pi\left(\theta_1 \mid \theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_d^{(j-1)}, \mathbf{y}\right)$;
    Draw $\theta_2^{(j)} \sim \pi\left(\theta_2 \mid \theta_1^{(j)}, \theta_3^{(j-1)}, \theta_4^{(j-1)}, \ldots, \theta_d^{(j-1)}, \mathbf{y}\right)$;
    Draw $\theta_3^{(j)} \sim \pi\left(\theta_3 \mid \theta_1^{(j)}, \theta_2^{(j)}, \theta_4^{(j-1)}, \theta_5^{(j-1)}, \ldots, \theta_d^{(j-1)}, \mathbf{y}\right)$;
    ...
    Draw $\theta_d^{(j)} \sim \pi\left(\theta_d \mid \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)}, \mathbf{y}\right)$;
end
```

When implementing the Gibbs sampler it is common to update one component at a time. However, it is possible to group relative parameters in blocks and update them jointly from their full conditional distribution given the data and remaining parameters.

1.2.2.2 The Metropolis-Hastings algorithm

Implementation of the Gibbs sampler requires the full conditional distribution of the components of $\boldsymbol{\theta}$ to be available in closed form. However, for many models this is not possible. An alternative approach is the general *Metropolis-Hastings* (MH) algorithm introduced by Hastings (1970). The method proceeds as shown in Algorithm 2.

Algorithm 2: The Metropolis-Hastings algorithm

```
Initialise $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)}\right)$;
for $j$ = 1, 2, ..., $J$ do
    Given $\boldsymbol{\theta}^{(j-1)}$, draw a candidate value $\boldsymbol{\theta}^*$ from the proposal density $q\left(\boldsymbol{\theta}^{(j-1)}, \cdot\right)$;
    Calculate $\alpha\left(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*\right) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^* \mid \mathbf{y})\, q\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j-1)}\right)}{\pi\left(\boldsymbol{\theta}^{(j-1)} \mid \mathbf{y}\right) q\left(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*\right)}\right\}$;
    Draw $u \sim \text{Uni}(0,1)$;
    if $u \leq \alpha\left(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*\right)$ then
        Set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^*$
    else
        Set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$
    end
end
```

There are several choices for the proposal distribution, the simplest being to allow q to depend only on its first argument that is, $q\left(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*\right) = q(\boldsymbol{\theta}^*)$. This leads to the *independence sampler*. Another popular choice is to select a proposal of the form $q\left(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*\right) = q\left(\left|\boldsymbol{\theta}^{(j-1)} - \boldsymbol{\theta}^*\right|\right)$ which is known as the *symmetric random walk Metropolis* and was first introduced by Metropolis *et al.* (1953). In the one-dimensional case, q is usually set to be a $\mathcal{N}\left(\theta^{(j-1)}, \sigma^2\right)$. The choice of the proposal distribution is crucial as it affects the performance of the Metropolis-Hastings algorithm. For example, setting the variance σ^2 of the Normal random walk Metropolis proposal too high may lead to very few proposed values being accepted, whereas setting it too low might result into high acceptance rate and hence slow convergence of the chain. For this reason there have been many attempts to develop adaptive algorithms automatically tune the proposal distribution, see for example Haario *et al.* (2001) and Roberts and Rosenthal (2009).

The MH algorithm can be used in conjunction with the Gibbs sampler for updating components for which the full conditionals are not available. This approach is known as the *Metropolis within Gibbs* algorithm.

1.2.2.3 Data augmentation

Data augmentation is broadly used when performing Bayesian analysis with unobserved data or latent variables. In such cases, it may be hard to integrate out the missing data $\mathbf{x}$ and therefore it is often more convenient to augment the parameter space by including $\mathbf{x}$, and use the joint posterior $\pi(\boldsymbol{\theta}, \mathbf{x} \mid \mathbf{y})$ as the target distribution. The reason is that one can then apply a two stage Gibbs sampler alternating between simulations of $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y})$ and $\mathbf{x}$ from $\pi(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})$, which are more tractable than simulating from $\pi(\boldsymbol{\theta} \mid \mathbf{y})$.

1.2.2.4 Practical implementations

The theory of MCMC guarantees convergence to the correct target distribution but the rate of convergence cannot be typically known in advance. Therefore, in practical applications it is advisable to examine the output of MCMC in order to check whether the chains have reached their stationary distribution. Convergence can be assessed by visual inspection of traceplots or using existing formal diagnostic tests which include Gelman and Rubin (1992), Geweke (1992) and Raftery and Lewis (1992). Moreover, in order to ensure that the samples taken are representative of the target posterior, the early values in the chain are usually discarded as a *burn-in.* The length of the burn-in generally depends on the starting values since it will take more iterations to reach stationarity when the initial state of the algorithm is far from the posterior mode. Finally, a further issue concerning MCMC implementation is the autocorrelation within chains. If the output exhibits strong autocorrelation then the samples contain less information regarding the desired distribution compared to when being independent. Also, chains with high autocorrelation may require more iterations to sufficiently explore the parameter space. When dealing with highly correlated chains, one common practice is to do *thinning* that is save the output every k-th iteration. However, it must be noted that there should be a balance between the amount of thinning and the cost of sampling.

1.2.3 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC, Duane *et al.*, 1987; Neal, 2011) mimics the evolution over time of a Hamiltonian system that is characterised by its position $(\mathbf{q})$

and momentum ($\mathbf{p}$). For the purposes of simulating from a posterior distribution of interest, $\mathbf{q} = \boldsymbol{\theta}$ and $\mathbf{p}$ is introduced artificially from a $\mathcal{N}_d(\mathbf{0}, \mathbf{M})$ distribution. Samples for $\boldsymbol{\theta}$ are obtained by simulating the dynamics of the system which are described by Hamilton's set of differential equations. However, solving Hamilton's equations is not possible in most practical applications and therefore integration is typically done with the leapfrog integrator which for stepsize ϵ, updates the current state $(\boldsymbol{\theta}(t), \mathbf{p}(t))$ to a new state $(\boldsymbol{\theta}(t+\epsilon), \mathbf{p}(t+\epsilon))$ as follows:

$$\begin{aligned}
\mathbf{p}\left(t + \frac{\epsilon}{2}\right) &= \mathbf{p}\,(t) + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}} \log \pi\,(\boldsymbol{\theta}(t) \mid \mathbf{y}) \\
\boldsymbol{\theta}\,(t+\epsilon) &= \boldsymbol{\theta}\,(t) + \epsilon \mathbf{M}^{-1}\mathbf{p}\left(t + \frac{\epsilon}{2}\right) \\
\mathbf{p}\,(t+\epsilon) &= \mathbf{p}\left(t + \frac{\epsilon}{2}\right) + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}} \log \pi\,(\boldsymbol{\theta}(t+\epsilon) \mid \mathbf{y}).
\end{aligned}$$

Note that an accept/reject step needs to be introduced at the end of integration in order to account for the error introduced by the discretisation. Usually, leapfrog integration is repeated for several number of steps L; the special case where $L = 1$ corresponds to the Metropolis-adjusted Langevin algorithm (Roberts and Rosenthal, 1998).

HMC gains efficiency by using gradient information from the target posterior and can outperform many MCMC algorithms under various scenarios, particularly in high dimensional problems (Neal, 2011; Girolami and Calderhead, 2011). Implementation of the method may be challenging since it requires the specification of the parameters ϵ, L and $\mathbf{M}$ but there has been work showing how these can be tuned automatically, see for example Hoffman and Gelman (2014).

1.2.4 Bayesian model selection

1.2.4.1 Bayes factor

The traditional approach to Bayesian model selection is concerned with the following situation. Suppose that the observed data $\mathbf{y}$ have been generated by some model $\mathcal{M}_k \in \mathcal{M}$, where $\mathcal{M}$ is a finite or countable set of competing models indexed by a parameter $k \in \mathcal{K}$. Each model $\mathcal{M}_k$ has it own vector of unknown parameters $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k$ of dimension d_k, where d_k may vary from model to model. Let $\pi(k)$ be the prior probability of model $\mathcal{M}_k$ such that $\sum_{k \in \mathcal{K}} \pi(k) = 1$. A choice between two models, say $\mathcal{M}_k$ and $\mathcal{M}_r$ $(k \neq r)$, often is based on the Bayes factor (Kass and

Raftery, 1995) defined as the ratio of posterior to prior odds in favour of model $\mathcal{M}_k$:

$$B_{kr} = \frac{\pi(k \mid \mathbf{y})/\pi(r \mid \mathbf{y})}{\pi(k)/\pi(r)}, \tag{1.2}$$

where $\pi(k \mid \mathbf{y})$ is the posterior probability of model $\mathcal{M}_k$ given by:

$$\pi(k \mid \mathbf{y}) = \frac{\pi(k)\,\pi(\mathbf{y} \mid k)}{\sum\limits_{r \in \mathcal{K}} \pi(r)\,\pi(\mathbf{y} \mid r)}. \tag{1.3}$$

In the above equation $\pi(\mathbf{y} \mid k)$ is the marginal likelihood of model $\mathcal{M}_k$ obtained as:

$$\pi(\mathbf{y} \mid k) = \int_{\boldsymbol{\theta}_k} \pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\,\pi(\boldsymbol{\theta}_k \mid k)\,\mathrm{d}\boldsymbol{\theta}_k.$$

where $\pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)$ is the likelihood function and $\pi(\boldsymbol{\theta}_k \mid k)$ is the model-specific prior distribution of $\boldsymbol{\theta}_k$. Combining Equations (1.2) and (1.3) we get,

$$B_{kr} = \frac{\pi(k \mid \mathbf{y})/\pi(r \mid \mathbf{y})}{\pi(k)/\pi(r)} = \frac{\pi(\mathbf{y} \mid k)}{\pi(\mathbf{y} \mid r)},$$

showing that the marginal likelihoods could also be considered as the quantities of key interest. Bayes factor may be interpreted as a measure of the evidence provided by the data in favour of $\mathcal{M}_k$ relative to $\mathcal{M}_r$; values for the Bayes factor greater than one support $\mathcal{M}_k$, whereas a Bayes factor less than one supports $\mathcal{M}_r$.

The integral involved in the computation of the Bayes factor can be evaluated analytically only in specific examples and therefore in any other case, one needs to employ asymptotic approximations or other computational methods. However, the standard MCMC algorithms cannot be applied because moves from one model to another involve changes in the dimension of the parameter space and thus generalised MCMC algorithms are required. Existing methodologies are based either on the estimation of the marginal likelihoods or equivalently on the estimation of the posterior probabilities for the competing models. In the following sections, we present some of the algorithms that are widely employed, focusing on the ones that are more relevant to our application.

1.2.4.2 Importance sampling

The problem of estimating the marginal likelihood of a model $\mathcal{M}_k$ can be re-written as:

$$\begin{aligned}\pi(\mathbf{y} \mid k) &= \int_{\boldsymbol{\theta}_k} \pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\, \frac{\pi(\boldsymbol{\theta}_k \mid k)}{q(\boldsymbol{\theta}_k)}\, q(\boldsymbol{\theta}_k)\, \mathrm{d}\boldsymbol{\theta}_k \\ &= \mathbb{E}_{q(\boldsymbol{\theta}_k)}\left[\frac{\pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\, \pi(\boldsymbol{\theta}_k \mid k)}{q(\boldsymbol{\theta}_k)}\right],\end{aligned}$$

where $q(\boldsymbol{\theta}_k)$ is called the *importance density*. Hence, an unbiased *importance sampling* (IS) estimator of the marginal likelihood can be obtain as (Ripley, 1987):

$$\widehat{P}_{IS}(\mathbf{y} \mid k) = \frac{1}{N}\sum_{i=1}^{N} \pi\left(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k^{(i)}\right) \frac{\pi\left(\boldsymbol{\theta}_k^{(i)} \mid k\right)}{q\left(\boldsymbol{\theta}_k^{(i)}\right)},$$

where $\boldsymbol{\theta}_k^{(i)}$ are *independent and identically distributed* (iid) samples from the proposal density $q(\boldsymbol{\theta}_k)$. The efficiency of the estimator depends on how well $q(\boldsymbol{\theta}_k)$ approximates the true posterior. Further, it needs to be ensured that $q(\boldsymbol{\theta}_k)$ has heavier tails compared to the unnormalised posterior $\pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\, \pi(\boldsymbol{\theta}_k \mid k)$ in order for the variance of the importance sampling estimator to be finite (Frühwirth-Schnatter, 2006).

1.2.4.3 Harmonic mean

An alternative way to write the marginal likelihood is:

$$\begin{aligned}\pi(\mathbf{y} \mid k) &= \left\{\int_{\boldsymbol{\theta}_k} \frac{q(\boldsymbol{\theta}_k)}{\pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\, \pi(\boldsymbol{\theta}_k \mid k)}\, \pi(\boldsymbol{\theta}_k \mid \mathbf{y}, k)\, \mathrm{d}\boldsymbol{\theta}_k\right\}^{-1} \\ &= \left\{\mathbb{E}_{\pi(\boldsymbol{\theta}_k \mid \mathbf{y}, k)}\left[\frac{q(\boldsymbol{\theta}_k)}{\pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\, \pi(\boldsymbol{\theta}_k \mid k)}\right]\right\}^{-1}.\end{aligned}$$

By taking $q(\boldsymbol{\theta}_k) = \pi(\boldsymbol{\theta}_k \mid k)$ leads to the *harmonic mean* (HM) estimator (Newton and Raftery, 1994):

$$\widehat{P}_{HM}(\mathbf{y} \mid k) = \left[\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\pi\left(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k^{(i)}\right)}\right]^{-1},$$

where $\boldsymbol{\theta}_k^{(i)}, i = 1, 2, \ldots, N$, are draws from the posterior distribution. This estimator is widely used because it can be directly calculated from the output of an MCMC algorithm. However, the harmonic mean estimator is known to exhibit large or even infinite variance for some models (Newton and Raftery, 1994).

1.2.4.4 Bridge sampling

Meng and Wong (1996) introduced the *bridge sampling* (BS) approach for computing ratios of normalising constants. The key identity for bridge sampling is,

$$\pi(\mathbf{y} \mid k) = \frac{\mathbb{E}_{q(\boldsymbol{\theta}_k)}\Big[\alpha(\boldsymbol{\theta}_k)\pi^*(\boldsymbol{\theta}_k \mid \mathbf{y}, k)\Big]}{\mathbb{E}_{\pi(\boldsymbol{\theta}_k|\mathbf{y},k)}\Big[\alpha(\boldsymbol{\theta}_k)q(\boldsymbol{\theta}_k)\Big]},$$

where $\alpha(\boldsymbol{\theta}_k)$ is an arbitrary function, $\pi^*(\boldsymbol{\theta}_k \mid \mathbf{y}, k)$ is the unnormalised posterior and $q(\boldsymbol{\theta}_k)$ is a normalised density. The formula leads to the bridge sampling estimator:

$$\widehat{P}_{BS}(\mathbf{y} \mid k) = \frac{\frac{1}{L}\sum_{\ell=1}^{L} \alpha\left(\tilde{\boldsymbol{\theta}}_k^{(\ell)}\right)\pi^*\left(\tilde{\boldsymbol{\theta}}_k^{(\ell)} \mid \mathbf{y}, k\right)}{\frac{1}{N}\sum_{i=1}^{N} \alpha\left(\hat{\boldsymbol{\theta}}_k^{(i)}\right) q\left(\hat{\boldsymbol{\theta}}_k^{(i)}\right)},$$

where $\tilde{\boldsymbol{\theta}}_k^{(1)}, \tilde{\boldsymbol{\theta}}_k^{(2)}, \ldots, \tilde{\boldsymbol{\theta}}_k^{(L)}$ are iid samples from $q(\boldsymbol{\theta}_k)$, and $\hat{\boldsymbol{\theta}}_k^{(1)}, \hat{\boldsymbol{\theta}}_k^{(2)}, \ldots, \hat{\boldsymbol{\theta}}_k^{(N)}$ are MCMC draws from the posterior, $\pi(\boldsymbol{\theta}_k \mid \mathbf{y}, k)$. The authors show that an asymptotically optimal choice of $\alpha(\boldsymbol{\theta}_k)$ in terms of expected relative error can be obtained using iid draws from both $q(\boldsymbol{\theta}_k)$ and $\pi(\boldsymbol{\theta}_k \mid \mathbf{y}, k)$,

$$\alpha(\boldsymbol{\theta}_k) \propto \frac{1}{L\,q(\boldsymbol{\theta}_k) + N\pi(\boldsymbol{\theta}_k \mid \mathbf{y}, k)}.$$

Meng and Wong (1996) suggest the following iterative procedure to calculate the bridge sampling estimator of the marginal likelihood:

$$\widehat{P}_{\mathrm{BS}}^{(t)}(\mathbf{y} \mid k) = \widehat{P}_{\mathrm{BS}}^{(t-1)}(\mathbf{y} \mid k)\; \frac{\frac{1}{L}\sum_{\ell=1}^{L} \frac{\tilde{\pi}^{(t-1)}\left(\tilde{\boldsymbol{\theta}}_k^{(\ell)} \mid \mathbf{y}, k\right)}{L\,q\left(\tilde{\boldsymbol{\theta}}_k^{(\ell)}\right) + N\,\tilde{\pi}^{(t-1)}\left(\tilde{\boldsymbol{\theta}}_k^{(\ell)} \mid \mathbf{y}, k\right)}}{\frac{1}{N}\sum_{i=1}^{N} \frac{q\left(\hat{\boldsymbol{\theta}}_k^{(i)}\right)}{L\,q\left(\hat{\boldsymbol{\theta}}_k^{(i)}\right) + N\,\tilde{\pi}^{(t-1)}\left(\hat{\boldsymbol{\theta}}_k^{(i)} \mid \mathbf{y}, k\right)}},$$

where $\tilde{\pi}^{(t-1)}(\boldsymbol{\theta}_k \mid \mathbf{y}, k) = \pi^*(\boldsymbol{\theta}_k \mid \mathbf{y}, k)/\widehat{P}_{\text{BS}}^{(t-1)}(\mathbf{y} \mid k)$. The recursion is repeated until convergence, using some other marginal likelihood estimator, e.g. importance sampling, to find the initial value $\widehat{P}_{\text{BS}}^{(0)}(\mathbf{y} \mid k)$.

1.2.4.5 Chib's method

Chib's method (Chib, 1995) is based on a rearrangement of Bayes' theorem:

$$\pi(\mathbf{y} \mid k) = \frac{\pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\, \pi(\boldsymbol{\theta}_k \mid k)}{\pi(\boldsymbol{\theta}_k \mid \mathbf{y}, k)},$$

which holds for all $\boldsymbol{\theta}_k$ in the support of the posterior $\pi(\boldsymbol{\theta}_k \mid \mathbf{y}, k)$. Therefore, for a fixed $\boldsymbol{\theta}_k = \boldsymbol{\theta}_k^*$ the log marginal likelihood can be estimated using:

$$\widehat{P}_{\text{Chib}}(\mathbf{y} \mid k) = \frac{\pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k^*)\, \pi(\boldsymbol{\theta}_k^* \mid k)}{\widehat{\pi}(\boldsymbol{\theta}_k^* \mid \mathbf{y}, k)}.$$

Even though the estimator is valid for any $\boldsymbol{\theta}_k^*$, such that $\pi(\boldsymbol{\theta}_k^* \mid \mathbf{y}, k) > 0$, it is more efficient to choose a point of high posterior density. As suggested by the author, estimation of the posterior ordinate $\pi(\boldsymbol{\theta}_k^* \mid \mathbf{y}, k)$ can be achieved by breaking the parameter vector into $D_k \leq d_k$ blocks:

$$\begin{aligned}\pi(\boldsymbol{\theta}_k^* \mid \mathbf{y},\, k) = \pi(\boldsymbol{\theta}_{k,1}^* \mid \mathbf{y},\, k) \times \pi(\boldsymbol{\theta}_{k,2}^* \mid \mathbf{y}, \boldsymbol{\theta}_{k,1}^*,\, k)& \\ \times \cdots \times \pi(\boldsymbol{\theta}_{k,D_k}^* \mid \mathbf{y}, \boldsymbol{\theta}_{k,1}^*, \boldsymbol{\theta}_{k,2}^*, \ldots, \boldsymbol{\theta}_{k,D_k-1}^*,\, k).&\end{aligned}$$

To calculate each term in this product, a separate MCMC is run in which only the unconditioned blocks of $\boldsymbol{\theta}_{k,j}^*$, $j = 1, 2, \ldots, D_k$, are updated and the appropriate remaining blocks are fixed at the high posterior density points. In Chib (1995), it is required that the full conditionals of each block are given in closed forms. Chib and Jeliazkov (2001) extended the methodology in order to allow for some of the blocks to be updated with MH steps.

1.2.4.6 Power posteriors

The *power posterior* (PP) approach to estimating the marginal likelihood (Friel and Pettitt, 2008) uses samples from the power posterior, defined as:

$$\pi(\boldsymbol{\theta}_k \mid \mathbf{y}, k, t) \propto \pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)^t\, \pi(\boldsymbol{\theta}_k \mid k),$$

where $t \in [0, 1]$ is a temperature parameter. Borrowing ideas from path sampling allows the log of the marginal likelihood to be represented in terms of the thermodynamic integral:

$$\log \pi(\mathbf{y} \mid k) = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}_k|\mathbf{y},k,t}\big[\log \pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\big]\, \mathrm{d}t,$$

where the expectation of the mean deviance is taken with respect to the power posterior at temperature t. The integral can be calculated numerically by discretising the temperature range as $0 = t_0 < t_1 < \ldots < t_n = 1$, and then the log marginal likelihood can be approximated by the trapezium rule as:

$$\log \widehat{P}_{PP}(\mathbf{x}) = \sum_{i=0}^{n-1}(t_{i+1} - t_i)\frac{\mathbb{E}_{\boldsymbol{\theta}_k|\mathbf{y},k,t_{i+1}}\big[\log \pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\big] + \mathbb{E}_{\boldsymbol{\theta}_k|\mathbf{y},k,t_i}\big[\log \pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\big]}{2}.$$

For each t_i, the expectation $\mathbb{E}_{\boldsymbol{\theta}_k|\mathbf{y},k,t_i}\big[\log \pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\big]$ are estimated using samples from the corresponding power posterior $\pi(\boldsymbol{\theta}_k \mid \mathbf{y}, k, t_i)$, which can be obtained with MCMC. The variability of the power posterior estimator depends on the total number and spacing of the t_i's. However, choosing a large number of temperatures requires considerably more computational effort. Finally, the precision of the estimate also depends on the total number of MCMC samples for each temperature.

1.2.4.7 Reversible jump MCMC algorithm

Reversible jump MCMC was introduced by Green (1995) as a generalisation of the MH algorithm. The method includes the model indicator k as an additional parameter and uses the joint posterior distribution,

$$\pi(k, \boldsymbol{\theta}_k \mid \mathbf{y}) = \frac{\pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\,\pi(k,\, \boldsymbol{\theta}_k)}{\displaystyle\sum_{k' \in \mathcal{K}} \int_{\boldsymbol{\theta}_{k'}} \pi(\mathbf{y} \mid k',\, \boldsymbol{\theta}_{k'})\pi(k',\, \boldsymbol{\theta}_{k'})\, \mathrm{d}\boldsymbol{\theta}_{k'}} \propto \pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\,\pi(\boldsymbol{\theta}_k \mid k)\,\pi(k),$$

as the target. The method can move between the candidate models as follows. Suppose that the current state of the Markov chain is $(k,\, \boldsymbol{\theta}_k)$; a move to a new state $(k',\, \boldsymbol{\theta}_{k'})$ is proposed by generating a random vector $\mathbf{u}$ from a proposal density q and setting $(k',\, \boldsymbol{\theta}_{k'}) = g_{k,k'}(k,\, \boldsymbol{\theta}_k, \mathbf{u})$, where $g_{k,k'}$ is some invertible function. The reverse move is implemented using a random vector $\mathbf{u}' \sim q'$ and setting $(k,\, \boldsymbol{\theta}_k) = g_{k',k}(k',\, \boldsymbol{\theta}_{k'}, \mathbf{u}')$, where $g_{k',k} = g^{-1}_{k,k'}$. Note that vectors $\mathbf{u}$ and $\mathbf{u}'$ play the role of matching the dimensions of $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k'}$, such that $d_k + d_{\mathbf{u}} = d_{k'} + d_{\mathbf{u}'}$. To achieve the correct limiting distribution, a proposed move from model k to model k' is accepted with probability $\alpha\big[(k, \boldsymbol{\theta}_k), (k', \boldsymbol{\theta}_{k'})\big] = \min\big(1, A_{kk'}\big)$, where $A_{kk'}$ is given by:

$$A_{kk'} = \frac{\pi(\mathbf{y} \mid k',\, \boldsymbol{\theta}_{k'})\, \pi(\boldsymbol{\theta}_{k'} \mid k')\, \pi(k')\, P_{k',k}\, q'(\mathbf{u}')}{\pi(\mathbf{y} \mid k,\, \boldsymbol{\theta}_k)\, \pi(\boldsymbol{\theta}_k \mid k)\, \pi(k)\, P_{k,k'}\, q(\mathbf{u})} \left| \frac{\partial(k',\, \boldsymbol{\theta}_{k'}, \mathbf{u}')}{\partial(k,\, \boldsymbol{\theta}_k, \mathbf{u})} \right|, \tag{1.4}$$

$P_{k,k'}$ being the probability of a move from model k to model k' and the final term is the Jacobian resulting from the transformation from $(k,\, \boldsymbol{\theta}_k, \mathbf{u})$ to $(k',\, \boldsymbol{\theta}_{k'}, \mathbf{u}')$. The reverse move proposal, from k' to k, is accepted with probability,

$$\alpha\big[(k', \boldsymbol{\theta}_{k'}), (k, \boldsymbol{\theta}_k)\big] = \min\big(1, A_{kk'}^{-1}\big).$$

In contrast to all previously described methods, RJMCMC does not approximate the marginal likelihoods but instead provides estimates of the posterior model probabilities. For each of the competing model, these are obtained as the proportion of iterations that the chain has spent in that model. The proposal densities q and q' must be well designed such that a sufficiently large proportion of transdimensional moves is being accepted. Note that for problems involving nested models, the standard practice is to set $d_{\mathbf{u}}$ or $d_{\mathbf{u}'}$ equal to zero, depending on which model has fewer parameters. Finally, when the proposed model k' is the same as the current model k we use a standard MCMC updates for the model parameters $\boldsymbol{\theta}_k$.

1.3 Hidden Markov models

Hidden Markov models (HMMs) consist of an unobserved state process of interest $\mathbf{X} = [X_t]_{t=1,2,\dots,T}$ and an process $\mathbf{Y} = [Y_t]_{t=1,2,\dots,T}$ of partial observations of $\mathbf{X}$ taken at successive time points $t = 1, 2, \dots, T$ (MacDonald and Zucchini, 1997). The former is a first order Markov chain and is typically assumed to have a discrete and finite state space $\mathcal{X}_s = \{0, 1, \dots, n_s\}$. The latter, can be any discrete or continuous process, possibly multivariate, depending on the application. Standard theory deals with *homogenous* HMMs, for which the transition probabilities of the hidden states P_{ij} from a state i to a new state j are constant over time and can be arranged in $(n_s + 1) \times (n_s + 1)$ transition matrix $\mathbf{P}$ with elements:

$$\mathbb{P}(X_t = j \mid X_{t-1} = i, \mathbf{X}_{1:t-2}) = \mathbb{P}(X_t = j \mid X_{t-1} = i) = P_{ij}, \quad \text{for } i, j \in \mathcal{X}_s,$$

where $t = 2, 3, \dots, T$, $\mathbf{X}_{1:t-2} = (X_1, X_2, \dots, X_{t-2})$. To complete the characterisation, we need to define the initial distribution of the hidden states at time $t = 1$, $\mathbb{P}(X_1 = i)$. Further, the model assumes that the distribution of any observation Y_t depends only on X_t and therefore given X_t, Y_t is conditionally independent of all remaining observations and hidden states. Figure 1.1 provides a graphical illustration of a hidden Markov model.

The standard HMM just described can be extended in several ways. One way to extend the model is to assume that the transition probabilities of the hidden states depend on the times which lead to the *non-homogenous* HMM. Another generalisation is the *coupled* hidden Markov model which is a collection of multiple HMMs of which the hidden states are coupled with some dependence structure. An example of an coupled hidden Markov model consisting of 2 interacting HMMs is shown in Figure 1.2. Finally, one can obtain a hidden semi-Markov model by assuming that the probability of a new state depends on the time already spent on the current state.

Figure 1.1: An illustration of a hidden Markov model showing the hidden states, x_t, and the observations, y_t.

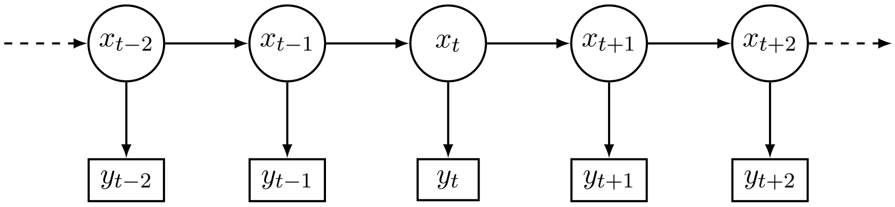
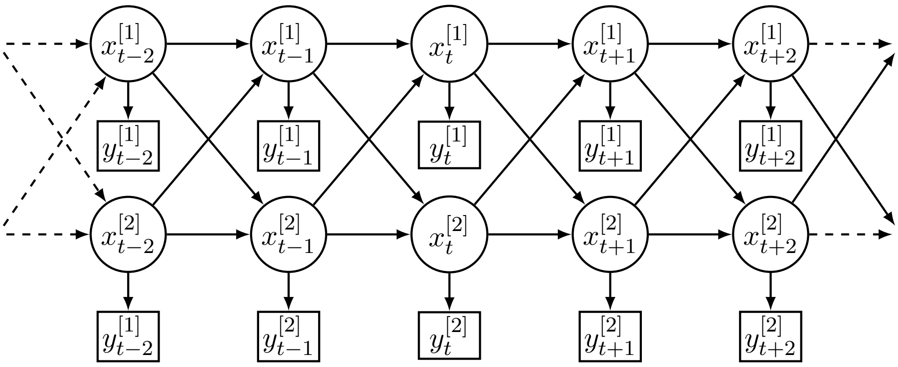


Figure 1.2: An illustration of a coupled hidden Markov model showing the hidden states, $x_t^{[c]}$, and the observations, $y_t^{[c]}$, for $c = 1, 2$.



1.4 Outline of the thesis

In this chapter, we have provided a brief background on current approaches for stochastic modelling of infectious diseases, emphasising problems where the underling epidemic process is only partially observed and have motivated the use of Markov chain Monte Carlo techniques, which we use as a tool for the methodology developed in this thesis.

The rest of the thesis is seeking to address some of the open problems described in Section 1.1.4 and is structured as follows. In Chapter 2 we perform an explanatory analysis of our datasets obtained from two independently conducted longitudinal studies of *Escherichia coli* (*E. coli*) O157:H7 in cattle. These datasets are used for our illustrations in subsequent work. Note that even though we use *E. coli* as an example, our methods can be easily applied to the analysis of numerous other infectious diseases.

Chapter 3 introduces an individual-based SIS model for the spread dynamics of an infectious disease among a population of individuals partitioned into households. The proposed hidden Markov model, that naturally accounts for partially observed data and imperfect test sensitivity, is used as the basic model for the methods developed throughout the thesis. Special attention is given to the data augmentation MCMC algorithm that is used to facilitate inferences for this model.

In Chapter 4 we consider the problem of Bayesian model selection in the presence of high-dimensional missing data, focusing on epidemiological applications where observations are gathered longitudinally and the population under investigation is organised in small groups. In particular, we outline an algorithm that combines ideas of MCMC, importance sampling and filtering to provide estimates of the marginal likelihood, and is well suited for small-scale epidemics. Even though several alternative approaches exist, there are currently only few studies assessing the performance of model selection methods in such settings. Hence, one of the main contributions of this chapter is the comparison of the proposed method with existing approaches, achieved through an extended simulation study on synthetic data generated in order to resemble real-life epidemiological problems. The importance of model selection procedures is further demonstrated in Chapter 5, where we successfully apply these methods to uncover new insights into the transmission dynamics of *E. coli* O157:H7 in cattle.

As discussed in Section 1.1.3, statistical inference for epidemic models often relies on data augmentation techniques for imputation of the hidden infection process. As a result, considerable progress has been made on developing such techniques, mainly using MCMC methods. However, as the dimensionality and complexity of the data increases some of these methods become inefficient, either because they produce chains with high autocorrelation or because they become computationally intractable. Motivated by this fact, in Chapter 6 we develop a novel MCMC algorithm, which is modification of the forward filtering backward sampling algorithm (Carter and Kohn, 1994), that achieves a good balance between computational complexity and mixing properties, and thus can be used to analyse epidemics on large populations. Even though our approach is developed under the assumption of a Markovian model, we show how this assumption can be relaxed leading to minor modifications in the algorithm. The performance of our method is assessed on both simulated and real data, considering models with simple structure but also complex dynamics, e.g. a model allowing for interactions between households.

The methodology developed in Chapter 6 permits us to extend the basic model in Chapter 7, in order to account for carriage of a disease with different serotypes. The growing availability of such data has lead to an increased demand for statistical tools that enable us to make use of this additional information. Our model addresses some of the limitations of the existing approaches, by simultaneously allowing for imperfect test sensitivities and serotype misclassification. The method is applied to a real dataset in order to further our understanding regarding the dynamics of various serotypes of *E. coli* O157:H7 in cattle, as well as to investigate between-serotype competition.

Finally, Chapter 8 summarises the contributions of the thesis and discusses some possible directions for future research.

Chapter 2

Motivating Escherichia Coli O157:H7 Datasets

2.1 Introduction

In this chapter, we present the two datasets that motivate our work and which will be analysed in the subsequent chapters of this thesis. The datasets are obtained from longitudinal studies regarding the presence of *E. coli* O157:H7 in cattle that were grouped in pens. The main difference between the two is that in the second dataset pens were sharing waterers, whereas in the first dataset no direct contact was possible between pens. *E. coli* O157:H7 is an important public health concern and cattle have been considered as the major animal reservoir of the pathogen (Ferens and Hovde, 2011; Wells *et al.*, 1991). It is therefore important to investigate the dynamics of *E. coli* O157:H7 transmission in cattle.

The chapter is organised as follows. In Section 2.2 we provide the details of experimental design as well as data collection for each dataset. This is followed by a preliminary analysis in Section 2.3, which focuses on aspects of the datasets such as the proportion of positive samples over the sampling period and associations between individuals. For dataset 1, there exists additional information regarding the serotypes in which the bacterium appears; this is presented in Section 2.4. Finally, in Section 2.5 we discuss our main findings and summarise the questions that arise.

2.2 Experimental data collection

In this section we present the datasets that motivate our analyses throughout the thesis. The first dataset is presented in Section 2.2.1 and the second dataset in Section 2.2.2.

2.2.1 Dataset 1

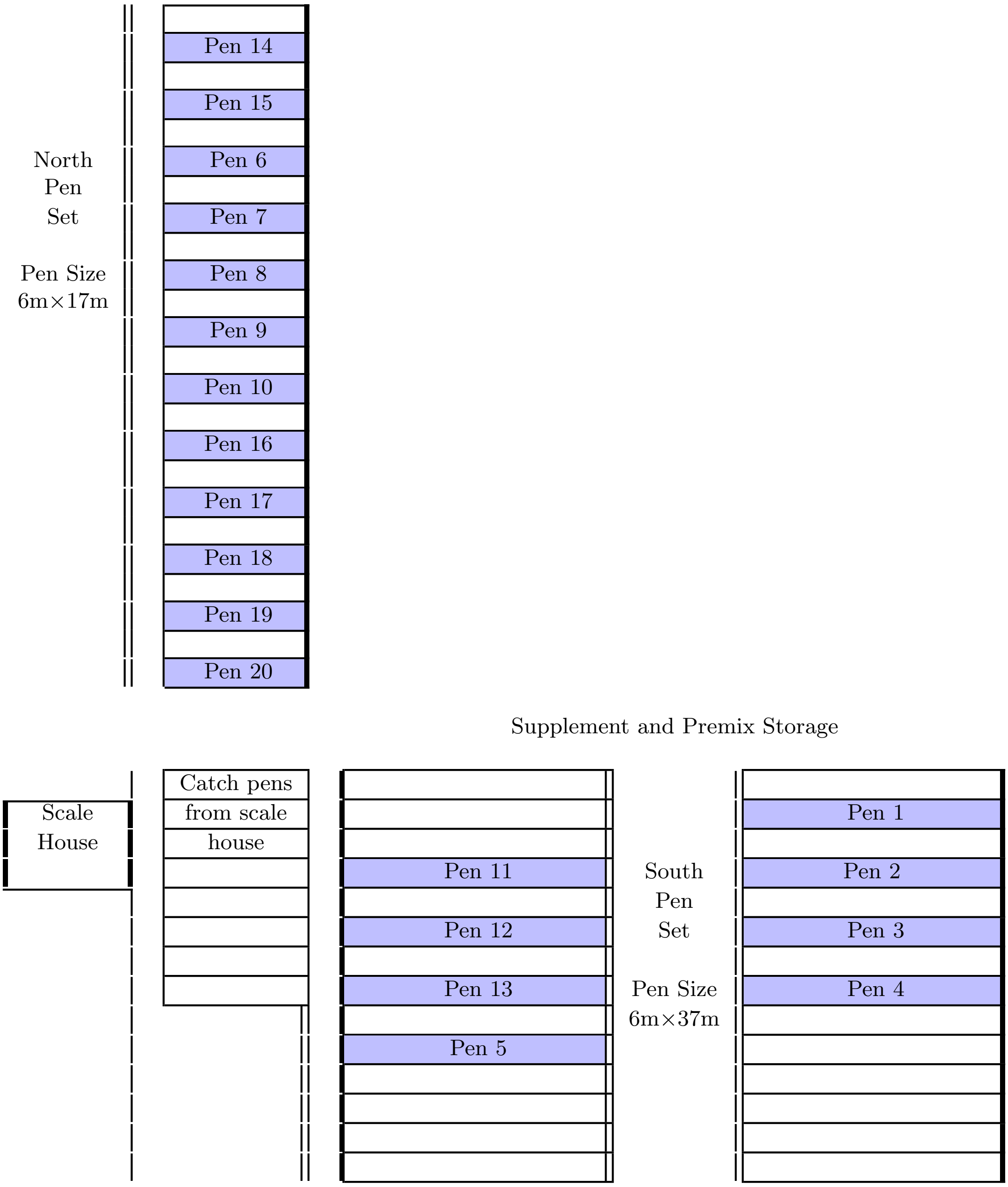
A longitudinal study of natural rectoanal junction (RAJ) colonisation and faecal excretion of *E. coli* O157:H7 was conducted in feedlot cattle (Cobbold *et al.*, 2007, for the full details). In this study 160 cattle, randomly assigned to twenty pens (eight animals each), were maintained and sampled within experimental research pens located at Canada on a working commercial feedlot with over 10,000 animals on the premises. Figure 2.1 is a schematic map with the locations of all the pens in this study. As can be seen, the pens were separated by an empty pen, ensuring that no direct contact was possible between animals of different pens. In addition, each pen had an individual water supply and a separate feed bunk. The animals were housed in North and South pens measuring 6m × 17m and 6m × 37m, respectively. The cattle included in the study were mixed breed, mixed sex and weighed approximately 350 kg.

Animals were sampled approximately twice per week, commencing approximately 3 days following pen assignment, over a 14-week period from 21 July to 27 October 2003. Briefly, at each sampling date two samples were collected from each animal: a recto-anal mucosal swab (RAMS) sample and a sample of freshly passed manure. The presence or absence of *E. coli* O157:H7 in each sample was determined by using Polymerase chain reaction (PCR). Therefore, the longitudinal data comprises of a set of result sequences from the two different tests, namely RAMS and faecal test. The test result of each individual was recorded at each day as 1, if the result was positive, 0 if it was negative and “NA” when either the sample was not taken or the animal withdrawn from the study before the completion, due to physical or behavioural problems in repeated handling. Figure A.1 in the Appendix is the graphical representation of the longitudinal data collected in *E. coli* O157:H7 dataset 1.

2.2.2 Dataset 2

Full details of the design and conditions of the experiment, as well as the collected data are described in details in Cernicchiaro *et al.* (2010), and we now summarise the salient points. Briefly, 168 Angus-cross beef steers, initially weighing 250 to 340 kg, were randomly allocated to a 24 pen beef feedlot research facility (seven animals each) located at the Ohio Agricultural Research and Development Center in Wooster, Ohio. All pens, constructed with metal gates and cables, had the same dimensions namely 5.4m × 5.4m. Pens 1-12 were placed adjacently from left to right, with the remaining 12 pens being exactly behind them counting at the opposite

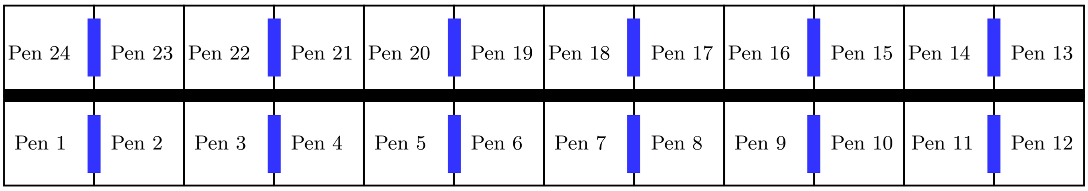
Figure 2.1: Canada Experimental Feedlot Pen Configuration. Bold lines along the feed alley represent concrete-based feed bunks. Double lines represent 20% porosity wind shelter. Coloured boxes indicates the pens used during the study period whereas white boxes represent empty pens.



direction, as we can see in Figure 2.2. Food was automatically distributed through the feed bunks, which were located between the two rows of pens. However, the orientation of the feed bunks did not allow any animal-to-animal contacts. In total 12 waterers were placed, and each one of these was shared between two adjacent pens, such that for example pens 1 and 2 had a common water supply. All the facilities, including waterers and feed bunks, were cleaned before the initiation of the experiment.

In the study, one RAMS and at least 10g of faecal grab samples were collected from each animal at 14-days interval, throughout a 22-week period from 21 November 2005 to 25 April 2006. Recovering the complete test results of the study was not possible; instead we observe whether an individual had been tested positive by at least one of the test which we denote by 1 or if both tests were negative which we denote by 0. Missing values were also recorded and we denote these by “NA”. A graphical representation of dataset 2 can be found in Figure A.2 of the Appendix.

Figure 2.2: Ohio agricultural feedlot research pen configuration. Waterers are shared between groups of two adjacent pens; blue coloured rectangles. Bold line between the two rows of pens represents the feed bunk.



2.3 Exploratory data analysis

In this section we perform a preliminary analysis, to gain some insight on the attributes of the data that we wish to examine in the investigations of the following chapters. The attributes that we study are listed below. For both datasets, we first summarise the proportion of positive outcomes among the test results. This information can be indicative of the true prevalence of the disease in the population of individuals considered. Further, for dataset 1 it may suggest some difference in the sensitivity of the 2 diagnostic tests considered.

Moreover, we investigate associations between cattle in the same pens as well as cattle in different pens. In particular, we use Kendall’s rank correlation coefficient (τ) to detect similarities in the test results of an individual compared to other individuals in the same pen but also individuals in different pens. For comparing the distribution of correlation coefficients within and among cattle pens we use non-parametric bootstrapping of the data (Efron, 1979), where R independent replicate datasets are obtained from the null hypothesis model and for each r-th realisation the appropriate test statistic value t^* is calculated, denoted by $t_1^*, t_2^*, \ldots, t_R^*$. A bootstrap p value corresponding to the test of alternative ($t^* > 0$) against the null ($t^* = 0$) may then be estimated by the proportion of bootstrap samples that yield a statistic greater than the observed statistic t_{obs} (Davison and Hinkley, 1997):

$$p_{\text{boot}} = \frac{1 + \#\{t_r^* \geq t_{obs}\}}{R + 1}.$$

A significance threshold of 0.05 is used for all hypothesis testing.

Finally, we study if there are more similarities in individuals housed in pens that are located close one to another in comparison to those further apart. To achieve so, we use Kendall rank correlation coefficient for pairwise comparisons and the log odds ratio for a pair of pens defined as:

$$\log\left(\frac{n_{00}\, n_{11}}{n_{01}\, n_{10}}\right),$$

where n_{ij} denotes the total number of days where an individual from the first pen was tested i while an individual from the second pen was tested as j, where $i, j \in \{0, 1\}$. Large values of the log odds ratio indicate strong association between pens and near zero values reveal little or no association. Note that for dataset 1, we define label 1 an individual that was found positive by at least one of the RAMS and faecal tests, similar to the information that is available for dataset 2.

2.3.1 Dataset 1

The picture of the collected data for individual cattle by sampling day is given in Figure A.1 in the Appendix. Note the presence of missing data in between sampling intervals; over the sampling period of 99 days samples were collected only on 27 days. Of the 20 pens that were tested, all pens have one or more samples identified as positive for *E. coli* O157:H7. Forty three cattle were tested negative by the RAMS test and fifty two by the faecal test throughout the 14-week study, even though at least one animal within the same pen was shedding the organism.

A total of 4266 pairs of faecal and RAMS samples were collected from all pens over the sampling period. A frequency table was generated from the subset of non missing data (Table 2.1). In agreement with previous reports (Gansheroff and

O'Brien, 2000; Low *et al.*, 2005), the majority of the samples were negative for *E. coli* O157:H7. Results from RAMS samples are in agreement with the faecal samples in 3948 (219 + 3729) of the 4266 samples (92.55%). It was also found that in 64 samples, the faecal sample tested positive and the RAMS sample tested negative, and the opposite occurred in 254 samples where the faecal sample tested negative and the RAMS sample was positive.

Table 2.1: Observed data from RAMS and faecal tests for the detection of *E. coli* O157:H7 in dataset 1, where tests stated as positive (1) or negative (0) according to the result of each individual animal.

		Faecal test		**Total**
		1	0	
RAMS test	1	219	254	473
	0	64	3729	3793
Total		283	3983	4266

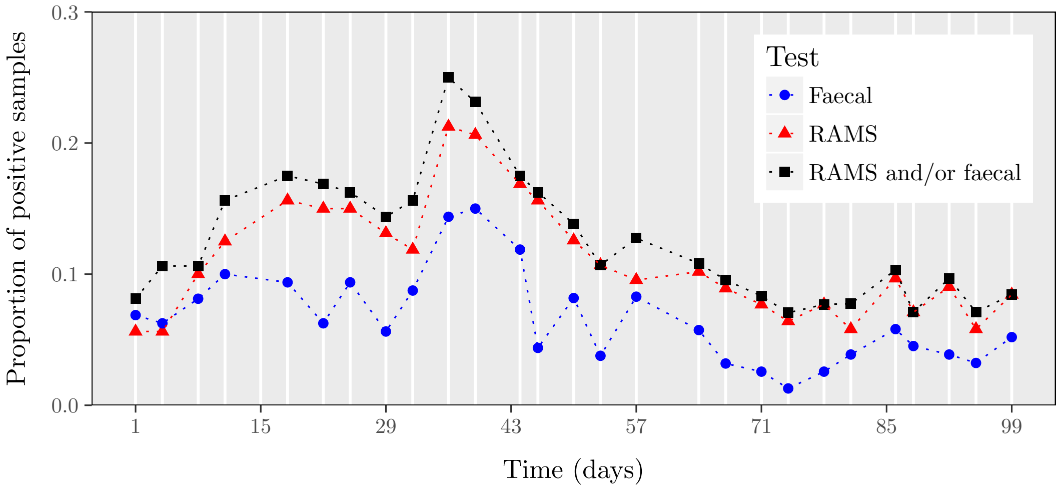
The proportion of *E. coli* O157:H7 positive tests for each sample date over the entire study period is presented in Figure 2.3. A pattern can be seen in the proportion of positive RAMS and/or faecal samples (black squares), whereby the first 6 weeks of the sampling period the proportion of positive samples increase, reaches the highest rate in day 36 and decreases during the last 8 weeks of the study. The same pattern can be observed for each test individually. However, we also note that the proportion of cattle detected shedding *E. coli* O157:H7 by the RAMS swabs is significantly higher than the proportion of positive faecal samples ($p < 0.001$, as determined by the Wilcoxon signed rank test). This result is consistent with previous reports (Greenquist *et al.*, 2005; Rice *et al.*, 2003) and may implies a better sensitivity for the RAMS test.

Difference in average within-pen correlation ($m_{\tau_{within}}$) and average between-pen correlation ($m_{\tau_{between}}$) is assessed with a bootstrap test using $t^* = m_{\tau_{within}} - m_{\tau_{between}}$ as a test statistic. The test is significant ($p = 0.002$), meaning that there is evidence that within-pen correlation is higher compared to between-pen correlation. The bootstrap distribution of the difference between the two means is shown in Figure 2.4, with the observed difference indicated by the black dot.

Testing $t^* = m_{\tau_{within}} = 0$ yields a p-value of < 0.001 suggesting that there is evidence of non-zero within-pen correlation, see also right panel of Figure 2.5. This fact is supported by visual inspection of the data in Figure A.1. For example, in pen number 15 positive samples are clustered within the interval 22-53. At the beginning of the study the samples collected from the animals are all tested negative and then, individual 3 began shedding a detectable level of the organism as demonstrated by the positive RAMS sample. This resulted in more positive samples for other cattle. Therefore, when modelling the spread of infection, it is important to capture dependencies among different animals living in the same pen.

In contrast, we find that the hypothesis $t^* = m_{\tau_{between}} = 0$ cannot be rejected with a p-value of 0.0645 suggesting that there is evidence of small between-pen correlation (see Figure 2.5, left panel). Plots of log odds ratios and correlations between all pairs of pens against centroid distance are displayed in Figure 2.6 and lead to a similar conclusion since they all lie close to zero. However, we observe a slight decrease of both log odds ratio and correlations for pens more 25 meters apart.

Figure 2.3: Proportion of RAMS swabs (triangles) and faeces samples (circles) that were positive for *E. coli* O157:H7 in feedlot cattle of dataset 1 over the sampling period. White vertical lines represent the days in which samples were taken.



2.3.2 Dataset 2

A picture of the whole study population is shown in Figure A.2 in the Appendix. We see that samples were taken in sparse time intervals of 2 weeks, thus representing only 12 days of the entire study duration of 156 days. Note that several individuals

Figure 2.4: Histogram of $t^* = m_{\tau_{whithin}} - m_{\tau_{between}}$ values from $R = 1999$ re-samples of *E. coli* O157:H7 dataset 1. The unshaded area of the histogram corresponds to the values of t^* larger than the observed value (black dot).

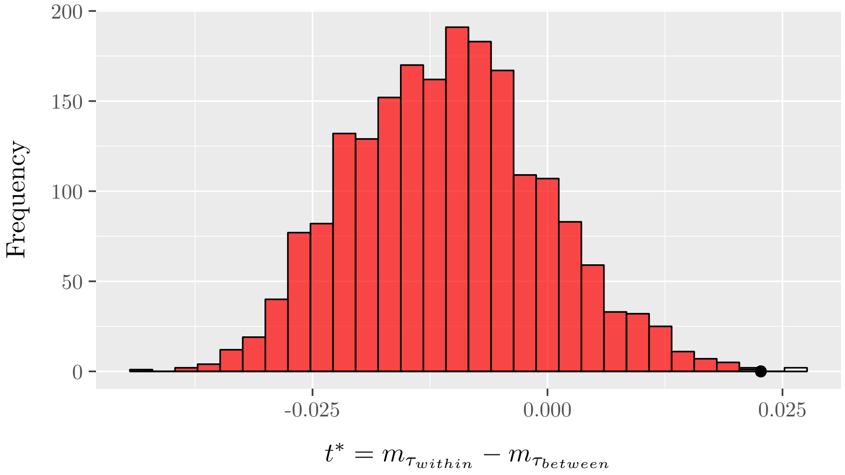
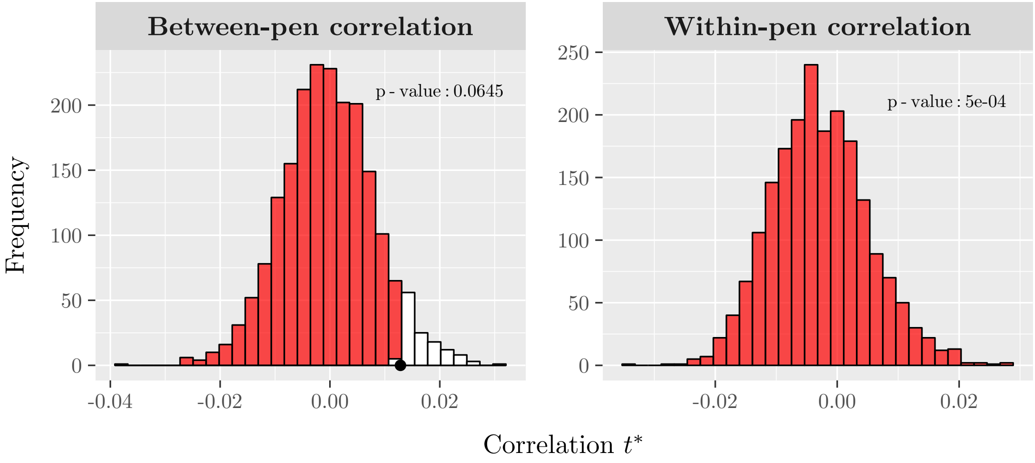
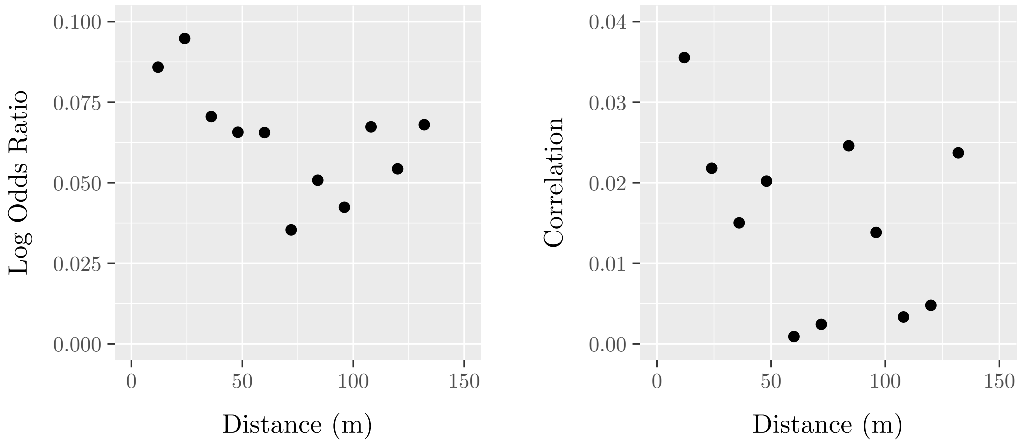


Figure 2.5: Histogram of correlation t^* values of within-pen (right panel) and between-pen (left panel) from $R = 1999$ re-samples of *E. coli* O157:H7 dataset 1. The unshaded area of the histogram corresponds to the values of t^* larger than the observed value (black dot).



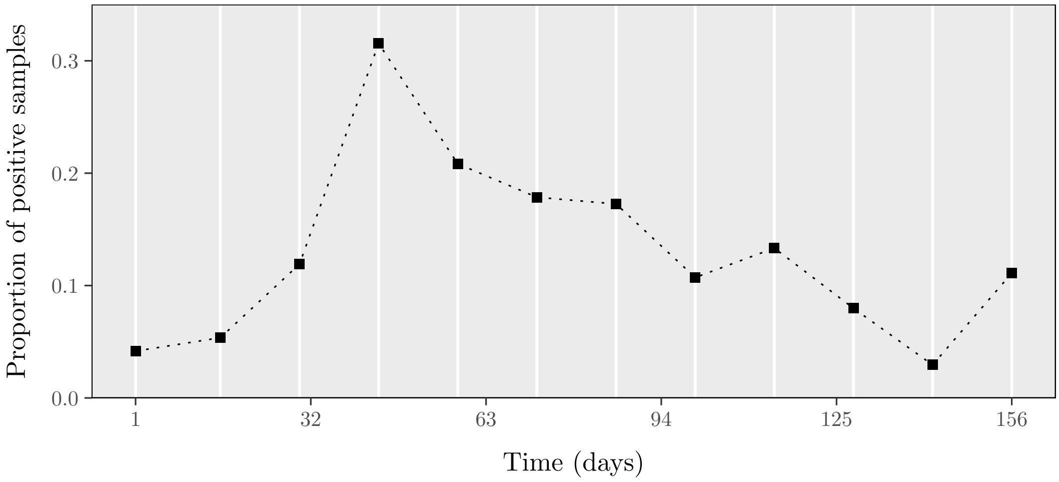
are withdrawn from the study before the completion, and in particular 70% (16/24) of the pens were not tested towards the end of the study. In total, 90% of the pre-scheduled tests were carried out. Of the 1828 samples collected, only 245 (13.40%) were found positive. Two pens, pens 5 and 22, have no occurrences of positive tests for any individual during the observation period. The proportion of positive test

Figure 2.6: Median (over all pens with the same distance) log odds ratios (left panel) and Kendall rank correlation coefficients (right panel) against distance, *E. coli* O157:H7 dataset 1.



outcomes over the study shows a similar pattern compared to dataset 1; in particular, we observe an increases of positive results around day 50 which gradually drops by the end of the study (Figure 2.7).

Figure 2.7: Proportion of RAMS swabs and/or faeces samples that were positive for *E. coli* O157:H7 in feedlot cattle of dataset 2 over the sampling period. Vertical white lines represent the observation days.



For this dataset, within-pen correlation is again found significant with bootstrap $p < 0.0001$ (right panel of Figure 2.8). Pen 18 (see Figure A.2) is indicative of the interactions of individuals within a pen. The first individual found positive is individual 2 at the third pre-scheduled sampling day and after that all individuals in the pen acquire the disease at some point within the following days.

Contrary to the first dataset, we found a non-zero between-pen correlation ($p = 0.003$, left panel of Figure 2.8). Moreover, the median log odds ratios for pairs of pens over centroid distance can be seen in the left panel of Figure 2.9. The figure shows a clear decrease in the log odds ratio with distance, with the largest log odds ratio corresponding to the smallest distance and the smallest log odds ratio corresponds to the largest distance. It is also noticeable that the magnitudes of the median log odds ratios in dataset 2 are considerably bigger than those in the dataset 1; only one (as compared with all) falls below 0.1. Similar pattern showing decrease in correlations with distance is observed in the right panel of Figure 2.9, with the five highest log odds ratios also have the highest correlation. These associations can be explained by the fact that several pens shared waterers and/or boundaries which potentially facilitate transmission of the disease. We formally test this hypothesis in Chapter 5.

FIGURE 2.8: Histogram of correlation t^* values of within-pen (right panel) and between-pen (left panel) from $R = 1999$ re-samples of *E. coli* O157:H7 dataset 2. The unshaded area of the histogram corresponds to the values of t^* larger than the observed value (black dot).

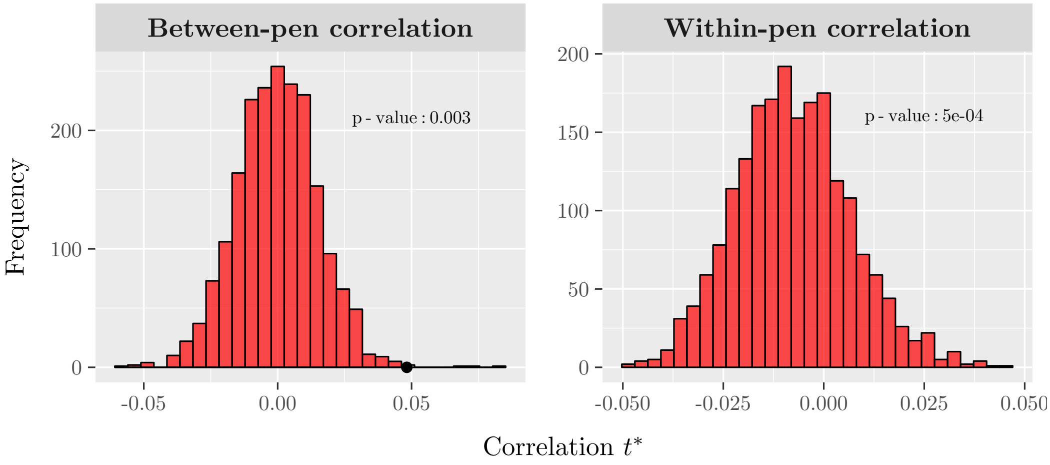
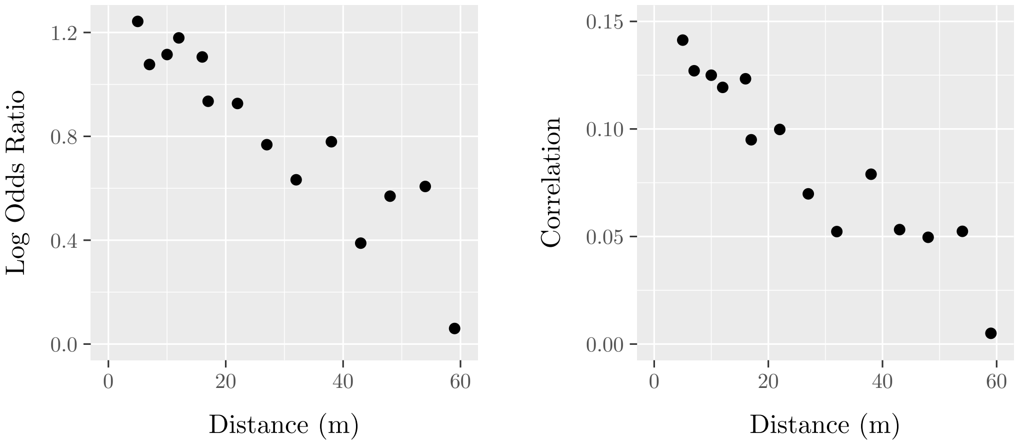


Figure 2.9: Median (over all pens with the same distance) log odds ratios (left panel) and Kendall rank correlation coefficients (right panel) against distance, *E. coli* O157:H7 dataset 2.



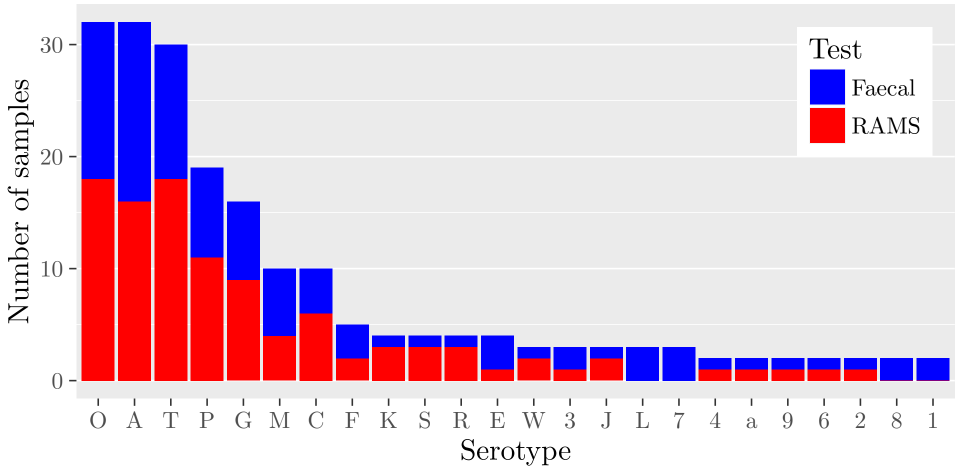
2.4 Serotype Data

Additional information for dataset 1 is available in the form of serotype data, comprising of a classification of the bacterium according to the structure of its isolate using pulsed-field gel electrophoresis (PFGE) as described by Tenover *et al.* (1997). More specifically, 12 positive samples (either RAMS or faecal) were randomly selected to be serotyped from each pen of the study. For 5 of the pens in the dataset the total number of serotyped samples varied from 5-11, either because less than 12 positive samples were obtained or because serotyping was unsuccessful. Overall, there were a total of 223 serotyped samples among the 756 positive samples, a proportion of almost 30%.

A total of 48 different serotypes were identified in the study population which we arbitrarily label according to the order in which they appeared in the PFGE typing. Of these, 24 appeared only once. Figure 2.10 illustrates the frequencies of serotypes, ranked from the most common to least common one, excluding the 24 serotypes that were unique. The following 7 serotypes represented the majority of positive samples: O, A, T, P, G, M and C.

Among the 160 cattle examined in the study, 106 (66.25%) gave at least one serotyped sample. For these, the median number of serotyped samples was 2 (min-max: 1-9). Figure 2.11 presents data collected in a subset of 4 pens. The data shown allows us to comment on the micro-epidemics of a serotype within a

Figure 2.10: Frequencies of serotypes identified in *E. coli* dataset 1, ranked from the most common to least common one, excluding the 24 serotypes that captured only once.



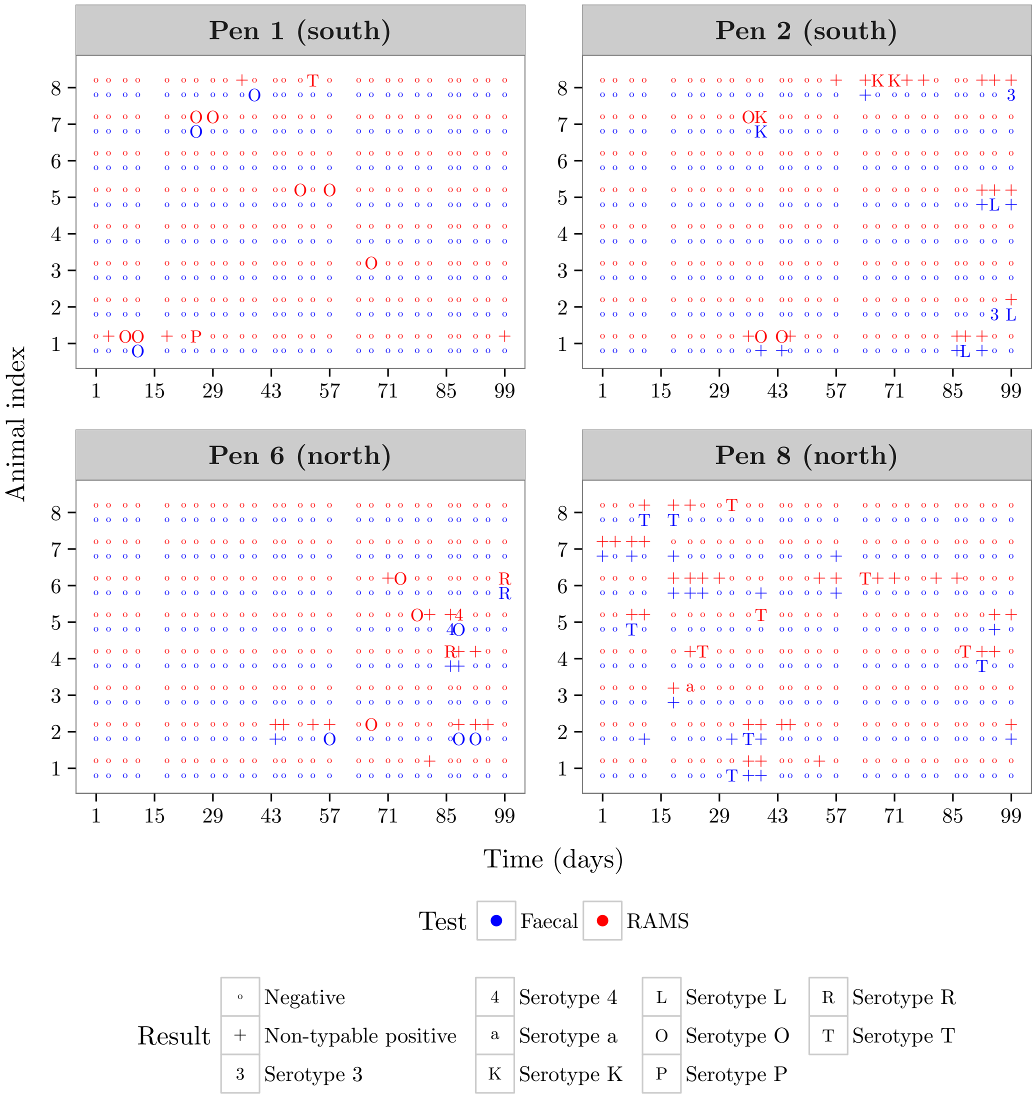
pen. For example, at the beginning of the study serotype T was detected in the samples collected from individual 5 in pen number 8, and then a micro-epidemic was observed with at least 5 individuals carrying serotype T during the following period. Note that, several individuals were never selected for serotyping (e.g. animal 7, pen 8).

Moreover, of the 223 serotyped samples, 22 pairs of positive samples were chosen to be serotyped, where as pair we define RAMS and faecal isolates from the same individual on the same sampling date. Of these, there were 19 occasions in which an animal was observed to carry the same serotype by RAMS and faecal isolates, and the remaining 4 were pairs of different serotypes (e.g. animal 5, pen 6 at day 88); this could be attributed to misclassification errors of the serotyping procedure, or it could be evidence of co-infection.

2.5 Discussion

In this chapter we have presented the datasets that motivate our analyses throughout the thesis. An exploratory look at the data revealed some interesting attributes that require further investigation. First of all, we found that our diagnostic tests often disagreed as to whether an individual was a carrier of the *E. coli* O157:H7 bacterium or not. This could be the consequence of imperfect tests and therefore needs to be accounted for. Further, strong correlations were detected for individuals belonging to the same pen. In particular, we saw that animals within a pen tended to give similar results on a given day of examination. Therefore, the development of a

FIGURE 2.11: RAMS and faecal samples (top red and bottom blue respectively) collected in pens 1, 2, 6 and 8 participating in the study. "∘" indicates negative sample, "+" indicates that the sample was positive but not chosen for serotyping; otherwise, serotype name is given.



model for these data should consider interaction between animals of the same pen. In Chapter 3 we build the basic model that describes the transmission of *E. coli* O157:H7 and incorporates the two previously mentioned characteristics.

For the second dataset, our analysis suggested positive between-pen correlation. In addition, we found that between-pen correlations were stronger for pens that were fewer metres apart. Notably, the highest associations arose from pens that shared either waters or boundaries. Hence, one could claim that one of these factors or both facilitate the transmission of the disease. However, formal assessment of such hypotheses requires more sophisticated tools compared to our approach in this chapter. In Chapter 4 we develop a framework which allows for hypothesis testing through model selection in the context of epidemiology and the proposed tools are applied to datasets 1 and 2 in Chapter 5.

Finally, PFGE analysis revealed several *E. coli* O157:H7 serotypes which possibly exhibit heterogeneity in transmissibility or the period for which they remain at the host. Another interesting question is whether there is competition between the subtypes. In Chapter 7 we propose a model for *E. coli* O157:H7 transmission that uses serotype information and also allows for serotype misclassification.

Chapter 3

A Non-Homogeneous Hidden Markov Model For Household Epidemic Data

3.1 Introduction

Escherichia coli O157:H7 is an important public health concern and it was first identified as a human pathogen in 1982 (Riley *et al.*, 1983). Infections in humans can result in diarrhea, haemorrhagic colitis, haemolytic uraemic syndrome, and even death (Teunis *et al.*, 2004; Karch *et al.*, 2005). Cattle have been considered as the major animal reservoir of the pathogen and play a significant role in the epidemics of human infection (as reviewed by Hussein and Sakuma (2005)). Human infections may arise from direct contact with cattle, indirectly via faecal material in the environment, contaminated food or other unknown sources (Mead *et al.*, 1997). Therefore, researchers have stressed that study of the disease in cattle is of vital importance in order to control the rate of the disease in humans (Rice *et al.*, 2003).

In general, understanding the dynamics of an infectious disease, that is the rate at which individuals acquire and recover from it, is crucial to control its spread. Longitudinal studies can be particularly informative for this aspect since they provide important insights into the transmission dynamics of a pathogen within a population. Household studies have also received considerable attention because they allow study of the dynamics of infection within a group of individuals as well as between groups within a community. However, estimating transmission parameters using this type of data presents numerous challenges.

A key facet of the problem is that the data are usually incomplete, in the sense that the times of acquiring and clearing infection are not directly observed. This is because individuals are often tested in sparse time intervals and the laboratory tests that are used to detect the disease are typically imperfect. In addition, epidemic studies are complicated because there are dependencies within the epidemic process (for example, the risk of acquiring infection might depend on the number of other colonised individuals) and much of the data can be missing (for example as a result of individual dropouts).

The aforementioned issues perplex statistical inferences since the evaluation of the likelihood involves summation over all possible infection states of individuals making this calculation highly involved. Therefore, many exact and approximate approaches have been developed, see O'Neill (2002) for an extended review. Several of the proposed methods consider the use of data imputation methods, in which the missing infection states are treated as additional model parameters. For example, Becker and Britton (1999) and Becker (1997) tackle the problem with an EM algorithm. An alternative approach is the use of MCMC methods which are currently popular techniques for analysing data on partially observed infectious diseases (Gibson and Renshaw, 1998; O'Neill and Roberts, 1999; Auranen *et al.*, 2000; Smith and Vounatsou, 2003; Cauchemez *et al.*, 2004; Streftaris and Gibson, 2004; Jewell *et al.*, 2009, for example).

Some of the literature focuses on hidden Markov models which assume that the observed process (the diagnostic test results) is associated with a hidden process (the true infection status) which itself can be described as a Markov chain (Bureau *et al.*, 2003; Smith and Vounatsou, 2003; Cooper and Lipsitch, 2004; Fearnhead and Meligkotsidou, 2004). Therefore, HMMs provide a natural framework to analyse infection dynamics in longitudinal studies where the observed data are subject to potential testing error due to poor sensitivity of the diagnostic used. Another advantage of this approach includes the ability to account for missing observations and testing intervals that are not equally spaced.

In relation to *E. coli* O157, previous mathematical modelling of the transmission dynamics includes both deterministic and stochastic frameworks (Turner *et al.*, 2003; Liu *et al.*, 2005; Matthews *et al.*, 2006a,b; Turner *et al.*, 2006, 2008; Ayscue *et al.*, 2009; Spencer *et al.*, 2015). However, the majority of these methodologies do not account for the fact that the detectability of *E. coli* O157 is not perfect. To overcome this limitation, we propose the use of a hidden Markov model to describe the transmission dynamics of *E. coli* O157 infection among a population of cattle which is partitioned into pens (households). The method is developed under the Bayesian framework which is facilitated by the use of the forward filtering backward sampling algorithm (Carter and Kohn, 1994) to impute the unobserved carriage states and hence infer epidemiological parameters relating to the duration and transmissibility of *E. coli* O157:H7 infection, as well as the test sensitivities. Our methodology provides an alternative to standard techniques for the study of

E. coli O157 transmission and is used as the basis for our studies in the following chapters.

The rest of this chapter is structured as follows. Section 3.2 contains details of the stochastic model for *E. coli* O157:H7 transmission in cattle, and in Section 3.3 we show the likelihood of this model. In Section 3.4 we develop the MCMC algorithm that is used for posterior simulations. Simulation studies are documented in Section 3.5. In Section 3.6 we apply the proposed method to the motivating datasets described in Section 2.2. Finally, we conclude with a discussion on limitations of the current model and possible extensions for further research in Section 3.7.

3.2 Model

We now introduce the basic model of *E. coli* O157:H7 transmission that is used throughout the thesis. We consider a population of individuals partitioned into pens of various sizes, so that each individual belongs to only one pen for entire period of the study. We employ a discrete time Susceptible-Infected-Susceptible model (Anderson and May, 1991) for the spread of infection in a pen. In the SIS model, each individual in the population is assumed to belong to one of two states for each day in the study: either susceptible or colonised (infected). Susceptible animals are those who do not have the disease and are able to be colonised by it; colonised animals have the disease and are able to infect susceptible animals.

Because infection with *E. coli* O157:H7 is usually harmless to cattle, we assume that individuals that are cleared from the disease can immediately re-acquire, without any period of immunity. More precisely, let $X_t^{[c,p]} \in \mathcal{X}_s = \{0, 1\}$ denote the true colonisation status of individual $c \in \{1, 2, \ldots, n_c^p\}$ in pen $p \in \{1, 2, \ldots, P\}$ at day $t \in \mathcal{T}^{c,p} = \{1, 2, \ldots, T^{c,p}\}$, where $X_t^{[c,p]} = 0$ represents the non-carriage state, $X_t^{[c,p]} = 1$ the colonised state and $T^{c,p}$ is the individual observation period of subject c in pen p. The individual observation period is defined as the time between the first sample taken from the subject and the last sample. For simplicity, we assume that the observation period is the same for all individuals within the same pen such that $\mathcal{T}^c = \{1, 2, \ldots, T^c\} \subseteq \{1, 2, \ldots, T\}$, starting from day 1 and T being the last day of the entire study. In Section B.2 of the Appendix we relax this assumption to account for dropouts.

Susceptible individuals can acquire colonisation via two possible routes: either from sources outside of the pen or through direct or indirect contact with other colonised individuals from inside the pen. Indirect infection within the pen occurs when a susceptible individual gets colonised through a source that was contaminated by a colonised individual such as faeces; any infection that cannot be attributed to a cause from within the pen is considered external transmission.

The probability that a susceptible individual is colonised from some external source on any given day is $1 - e^{-\alpha}$. Therefore, individuals avoid colonisation with a constant external colonisation probability $e^{-\alpha}$, which is equal to the probability of there being no events in a Poisson process with rate α. Hence, α can be viewed as the external colonisation rate. We further assume that at a given day t a susceptible individual has a probability $e^{-\beta X_t^{[c,\,p]}}$ of avoiding colonisation on day $t+1$ from an infected individual c in the pen; therefore the overall probability of avoiding colonisation from all colonised individuals within the pen at time t is $e^{-\beta \sum_{c=1}^{n_c^p} X_t^{[c,\,p]}}$.

Once an individual is colonised, they remain so for a number of days called the colonisation (infection) period. In this model, we assume that the duration of carriage has a Geometric distribution with mean m, that is common to all colonised individuals. Note that the Geometric distribution is the discrete-time analogue of the exponential distribution, which is typically used to model colonisation period in continuous-time stochastic epidemic models. Moreover, like its continuous analogue, the Geometric has the memoryless property, that is, the probability of leaving the colonised state is constant in time, and thus is a mathematically convenient choice for modelling the duration of colonisation.

Based on these assumptions, the actual hidden carriage process for each individual c in a given pen p, is modelled as a discrete-time two-state Markov process with transition probabilities between states defined as:

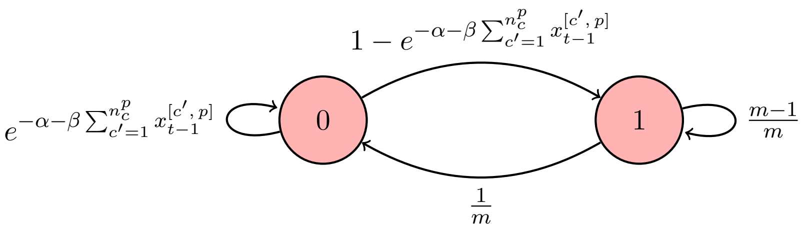
$$P^{[c,\,p]}_{x_{t-1}^{[c,\,p]},\,x_t^{[c,\,p]},\,t} := \mathbb{P}\left(X_t^{[c,\,p]} = x_t^{[c,\,p]} \mid X_{t-1}^{[c,\,p]} = x_{t-1}^{[c,\,p]},\, \mathbf{X}_{t-1}^{[-c,\,p]} = \mathbf{x}_{t-1}^{[-c,\,p]}, \alpha, \beta, m\right)$$

$$= \begin{cases} 1 - \exp\Big\{ -\alpha - \beta \sum_{c'=1}^{n_c^p} x_{t-1}^{[c',\,p]} \Big\} & \text{if } x_{t-1}^{[c,\,p]} = 0 \text{ and } x_t^{[c,\,p]} = 1, \\ \exp\Big\{ -\alpha - \beta \sum_{c'=1}^{n_c^p} x_{t-1}^{[c',\,p]} \Big\} & \text{if } x_{t-1}^{[c,\,p]} = 0 \text{ and } x_t^{[c,\,p]} = 0, \\ \dfrac{m-1}{m} & \text{if } x_{t-1}^{[c,\,p]} = 1 \text{ and } x_t^{[c,\,p]} = 1, \\ \dfrac{1}{m} & \text{if } x_{t-1}^{[c,\,p]} = 1 \text{ and } x_t^{[c,\,p]} = 0, \end{cases} \tag{3.1}$$

for $t = 2, 3, \ldots, T^p$, where $\alpha > 0$ and $\beta > 0$ denote the external and within-pen colonisation rates, respectively, and $m \geq 1$ denotes the mean colonisation period. Also, $\mathbf{X}_{t-1}^{[-c,\,p]}$ denotes the vector $\mathbf{X}_{t-1}^{[c,\,p]} = \left(X_{t-1}^{[1,\,p]}, X_{t-1}^{[2,\,p]}, \ldots, X_{t-1}^{[n_c^p,\,p]}\right)$ excluding $X_{t-1}^{[c,\,p]}$. The first and the last case in Equation (3.1) correspond to the colonisation ($0 \mapsto 1$) and clearance ($1 \mapsto 0$) probabilities, respectively. This parameterisation defines a non-homogeneous Markov model since it allows the probability of colonisation to depend on a sufficient statistic of the previous state of all individuals, namely the number of colonised individuals. Finally, since individuals were randomly assigned to pens, we assume that at the beginning of the study each animal is colonised independently with probability $\mathbb{P}\left(X_1^{[c,\,p]} = 1 \mid \nu\right) = \nu$. Figure 3.1 is a graphical representation of the individual-based SIS transmission model for a given pen.

However, as noted above, the underlying carriage process is not directly observed. Instead, for each individual we obtain the results of 2 diagnostic tests that are taken on pre-specified discrete times. Let $O^{c,p} \subseteq \mathcal{T}^p$ denote the set of prescheduled observations times of individual c in pen p, $c = 1, 2, \ldots, n_c^p$, $p = 1, 2, \ldots, P$ and let $U^{c,p} = \mathcal{T}^p \setminus O^{c,p}$ denote the times that the individual was not tested for the presence of the colonising organism. Therefore our model allows for the possibility that individuals may be tested on different days or not tested at all in some occasions. Such instances are very common in household epidemic data where the members of a family may not be tested on the same day. Nevertheless, in our applications animals were tested at common time points. Let $Y_t^{[c,\,p]} = \left(R_t^{[c,\,p]}, F_t^{[c,\,p]}\right)$ be the observed results, possibly misclassified, of the two diagnostic tests, $R_t^{[c,\,p]}$ for RAMS and $F_t^{[c,\,p]}$ for faecal. At each day $t \in O^{c,\,p}$ an individual c at pen p is classified as 1, if the test result is positive or 0, if we have a negative result. When $t \in U^{c,\,p}$ we have a missing value which we denote by "NA". We assume that the observed test results in each individual are independent Bernoulli variables, with the probability of a positive outcome depending on the underlying true colonisation status, as well as the probabilities of correctly or incorrectly classifying the observed outcome given

Figure 3.1: Possible routes of transitions between different states of an individual c in pen p. A susceptible animal (0) can become infected (1). Animals can recover to become susceptible again. The respective arrow annotations denote the transition probabilities.


$1 - e^{-\alpha-\beta\sum_{c'=1}^{n_c^p} x_{t-1}^{[c',\,p]}}$
$e^{-\alpha-\beta\sum_{c'=1}^{n_c^p} x_{t-1}^{[c',\,p]}}$
0
1
$\frac{m-1}{m}$
$\frac{1}{m}$

the true state. More specifically:

$$f^{R}_{x_t^{[c,\,p]}}\left(\mathrm{r}_t^{[c,\,p]} \mid \theta_R\right) := \mathbb{P}\left(R_t^{[c,\,p]} = \mathrm{r}_t^{[c,\,p]} \mid X_t^{[c,\,p]} = x_t^{[c,\,p]}, \theta_R\right)$$
$$= \begin{cases} \left(x_t^{[c,\,p]}\,\theta_R\right)^{\mathrm{r}_t^{[c,\,p]}} \left(1 - x_t^{[c,\,p]}\,\theta_R\right)^{1-\mathrm{r}_t^{[c,\,p]}} & \text{if } t \in O^{c,\,p}, \\ 1 & \text{if } t \in U^{c,\,p}, \end{cases} \tag{3.2}$$

$$f^{F}_{x_t^{[c,\,p]}}\left(\mathrm{f}_t^{[c,\,p]} \mid \theta_F\right) := \mathbb{P}\left(F_t^{[c,\,p]} = \mathrm{f}_t^{[c,\,p]} \mid X_t^{[c,\,p]} = x_t^{[c,\,p]}, \theta_F\right)$$
$$= \begin{cases} \left(x_t^{[c,\,p]}\,\theta_F\right)^{\mathrm{f}_t^{[c,\,p]}} \left(1 - x_t^{[c,\,p]}\,\theta_F\right)^{1-\mathrm{f}_t^{[c,\,p]}} & \text{if } t \in O^{c,\,p}, \\ 1 & \text{if } t \in U^{c,\,p}, \end{cases} \tag{3.3}$$

$$f_{x_t^{[c,\,p]}}\left(y_t^{[c,\,p]} \mid \theta_R, \theta_F\right) := f^{R}_{x_t^{[c,\,p]}}\left(\mathrm{r}_t^{[c,\,p]} \mid \theta_R\right) \times f^{F}_{x_t^{[c,\,p]}}\left(\mathrm{f}_t^{[c,\,p]} \mid \theta_F\right), \tag{3.4}$$

where $\theta_R = \mathbb{P}\left(R_t^{[c,\,p]} = 1 \mid X_t^{[c,\,p]} = 1\right)$ is the sensitivity of the RAMS test and $\theta_F = \mathbb{P}\left(F_t^{[c,\,p]} = 1 \mid X_t^{[c,\,p]} = 1\right)$ is the sensitivity of the faecal test. These probabilities are assumed to be the same for all individuals at all observation times. The formulation implies that the test specificities $\mathbb{P}\left(R_t^{[c,\,p]} = 0 \mid X_t^{[c,\,p]} = 0\right)$ and $\mathbb{P}\left(F_t^{[c,\,p]} = 0 \mid X_t^{[c,\,p]} = 0\right)$ are both 100%. Moreover, Equation (3.4) indicates that the two tests are statistically independent conditional on the true colonisation status of the subject. This assumption means that knowledge of the outcome of one diagnostic test does not provide any information about the outcome of other diagnostic tests, conditional on the true disease status (Toft *et al.*, 2005).

Note that by letting $\mathbf{X}_t^{[1:n_c^p,\,p]} = \left(X_t^{[1,\,p]}, X_t^{[2,\,p]}, \ldots, X_t^{[n_c^p,\,p]}\right)$ denote the vector of the hidden states of all individuals in pen p at given day t, we can convert the model to a single HMM for each pen, in which $\mathbf{X}_t^{[1:n_c^p,\,p]} \in \mathcal{X}_s^{n_c^p} = \{0,1\}^{n_c^p}$ denotes the state of the model at time t. Therefore, for a pen with n_c^p interacting individuals, the equivalent HMM will have a state space of size $2^{n_c^p}$ and its transition probabilities can be decomposed as:

$$\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,\,p]} = \mathbf{x}_t^{[1:n_c^p,\,p]} \mid \mathbf{X}_{t-1}^{[1:n_c^p,\,p]} = \mathbf{x}_{t-1}^{[1:n_c^p,\,p]}, \alpha, \beta, m\right)$$
$$= \prod_{c=1}^{n_c^p} \mathbb{P}\left(X_t^{[c,\,p]} = x_t^{[c,\,p]} \mid X_{t-1}^{[c,\,p]} = x_{t-1}^{[c,\,p]}, \mathbf{X}_{t-1}^{[-c,\,p]} = \mathbf{x}_{t-1}^{[-c,\,p]}, \alpha, \beta, m\right)$$
$$= \prod_{c=1}^{n_c^p} P^{[c,\,p]}_{x_{t-1}^{[c,\,p]},\, x_t^{[c,\,p]},\, t}, \qquad \text{for } t \in 2, 3, \ldots, T^p. \tag{3.5}$$

Moreover, the initial distribution on the state is assumed to factorise as:

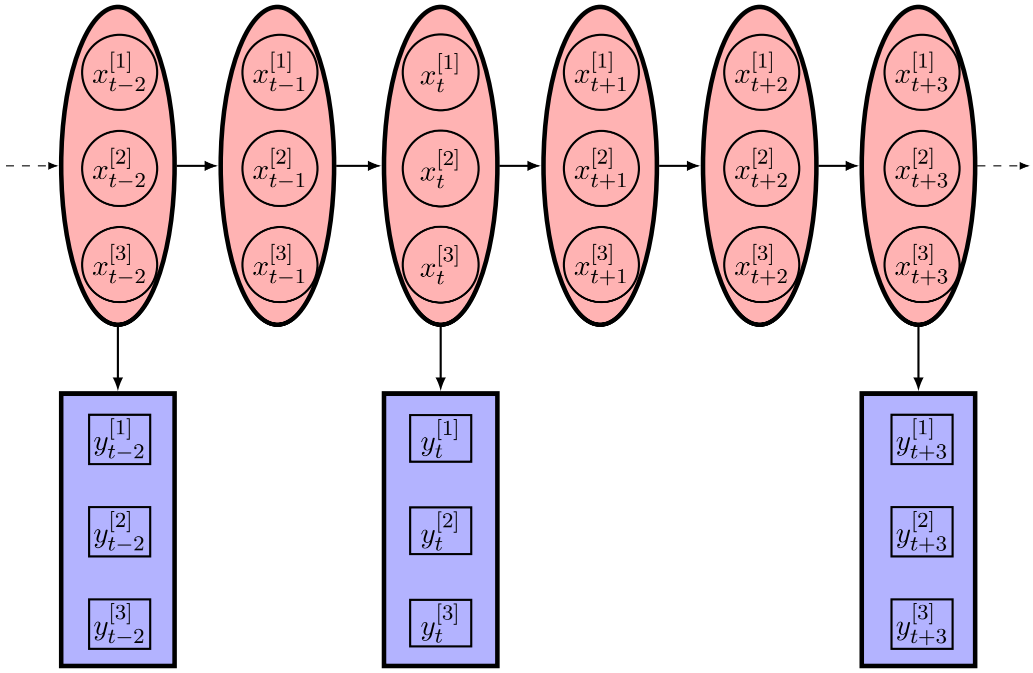
$$\mathbb{P}\left(\mathbf{X}_1^{[1:n_c^p,\,p]} = \mathbf{x}_1^{[1:n_c^p,\,p]} \mid \nu\right) = \prod_{c=1}^{n_c^p} \mathbb{P}\left(X_1^{[c,\,p]} = x_1^{[c,\,p]} \mid \nu\right) = \prod_{c=1}^{n_c^p} \nu^{\,x_1^{[c,\,p]}} (1-\nu)^{1-x_1^{[c,\,p]}}. \tag{3.6}$$

Finally, since the observation of each individual at any time is independent of other observations and states given the state of that subject at that time, we can write:

$$\begin{aligned} f_{\mathbf{x}_t^{[1:n_c^p,\,p]}}\left(\mathbf{y}_t^{[1:n_c^p,\,p]} \mid \theta_R, \theta_F\right) &:= \mathbb{P}\left(\mathbf{Y}_t^{[1:n_c^p,\,p]} \mid \mathbf{X}_t^{[1:n_c^p,\,p]} = \mathbf{x}_t^{[1:n_c^p,\,p]}, \theta_R, \theta_F\right) \\ &= \prod_{c=1}^{n_c^p} f_{x_t^{[c,\,p]}}\left(y_t^{[c,\,p]} \mid \theta_R, \theta_F\right). \end{aligned} \tag{3.7}$$

Therefore taken together Equations (3.5)–(3.7) constitute a non-homogenous hidden Markov model with multivariate hidden states $\mathbf{X}_t^{[1:n_c^p,\,p]}$ and multivariate observations $\mathbf{Y}_t^{[1:n_c^p,\,p]}$. An illustration of this HMM is given in Figure 3.2 for a pen with 3 individuals.

FIGURE 3.2: A hidden Markov model represented as a dynamic Bayesian network, with three individuals for a given pen ($n_c = 3$) and possibly several missing observations. Red nodes denote hidden states and blue nodes denote observations.



3.3 Likelihood computation

Let $\mathbf{X}^p = \left[X_t^{[c,\,p]}\right]_{c=1,2,\ldots,n_c^p;\, t=1,2,\ldots,T^p}$ denote the collection of hidden states for all individuals in pen p and $\mathbf{X}$ denote the collection of $\mathbf{X}^p$, $p = 1, 2, \ldots, P$. We define $\mathbf{R}^p, \mathbf{F}^p, \mathbf{R}, \mathbf{F}$ in a similar fashion so that $\mathbf{Y}^p = (\mathbf{R}^p, \mathbf{F}^p)$ is the observed data for pen p and $\mathbf{Y} = (\mathbf{R}, \mathbf{F})$ denotes the full observed data. For each pen p with T^p timepoints and n_c^p subjects, there are $2^{n_c^p \times T^p}$ possible status sequences. Let $\mathcal{X}_s^{n_c^p \times T^p} = \{0, 1\}^{n_c^p \times T^p}$ be the space of possible states taken by $\mathbf{X}^p$, $p = 1, 2, \ldots, P$. Therefore for pen p, the likelihood of the observed data given the model parameters can be written as:

$$\begin{aligned}
\pi(\mathbf{Y}^p \mid \boldsymbol{\theta}) &= \sum_{\boldsymbol{\omega}^p \,\in\, \mathcal{X}_s^{n_c^p \times T^p}} \mathbb{P}(\mathbf{Y}^p \mid \mathbf{X}^p = \boldsymbol{\omega}^p, \boldsymbol{\vartheta})\, \mathbb{P}(\mathbf{X}^p = \boldsymbol{\omega}^p \mid \boldsymbol{\phi}) \\
&= \sum_{\boldsymbol{\omega}^p \,\in\, \mathcal{X}_s^{n_c^p \times T^p}} \left[\prod_{c=1}^{n_c^p} \prod_{t=1}^{T^p} f_{\omega_t^{[c,\,p]}} \left(y_t^{[c,\,p]} \mid \theta_R, \theta_F \right) \right. \\
&\qquad \left. \times \prod_{c=1}^{n_c^p} \left(\mathbb{P}\left(X_1^{[c,\,p]} = \omega_1^{[c,\,p]} \mid \nu \right) \prod_{t=2}^{T^p} P^{[c,\,p]}_{\omega_{t-1}^{[c,\,p]}, \omega_t^{[c,\,p]}, t} \right) \right],
\end{aligned}$$

where $\boldsymbol{\phi} = (\alpha, \beta, m, \nu)$ denotes the vector of transmission parameters, $\boldsymbol{\vartheta} = (\theta_R, \theta_F)$ is the vector of observation parameters and $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\vartheta})$ denotes the complete model parameters.

Since we assume that pens are independent one of another, the likelihood of the full observation data, $\mathbf{Y}$ given $\boldsymbol{\theta}$, is given by the product of the individual pen likelihoods:

$$\pi(\mathbf{Y} \mid \boldsymbol{\theta}) = \prod_{p=1}^{P} \pi(\mathbf{Y}^p \mid \boldsymbol{\theta}). \tag{3.8}$$

Direct evaluation of Equation (3.8) involves summation over all possible colonisation states of individuals which in practice is computationally feasible only for very small pen sizes, total number of pens and short sampling periods. In most applications, including ours, calculation of the likelihood above is intractable in practice. The approach adopted in this work to overcome this difficulty is to use Bayesian inference, in particular MCMC techniques. The exact methodology is described in the following section.

3.4 Posterior sampling algorithm

We aim to make inferences on the model parameters given the observed data. The task is complicated by the fact that evaluation of the model likelihood is intractable since the true colonisation process is unobserved. We therefore consider this unobserved process as an additional parameter and estimate it along with the model parameters. Model fitting is performed within the Bayesian framework by imputing the unobserved colonisation states. Hence, our objective is to explore the joint posterior density of the augmented disease process and the model parameters conditional upon the observed test results, given by:

$$\pi(\mathbf{X}, \tilde{\alpha}, \tilde{\beta}, \tilde{m}, \nu, \theta_R, \theta_F \mid \mathbf{Y}) \propto \pi(\mathbf{Y} \mid \mathbf{X}, \theta_R, \theta_F)\ \pi(\mathbf{X} \mid \tilde{\alpha}, \tilde{\beta}, \tilde{m}, \nu)\, \pi(\tilde{\boldsymbol{\theta}}),$$

where $\pi(\tilde{\boldsymbol{\theta}})$ is the prior of the transformed model parameters $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\phi}}, \boldsymbol{\vartheta}) = (\tilde{\alpha}, \tilde{\beta}, \tilde{m}, \nu, \theta_R, \theta_F)$, $\tilde{\alpha} = \log(\alpha)$, $\tilde{\beta} = \log(\beta)$ and $\tilde{m} = m - 1$. The model parameters are assigned the following prior distributions, independent one of another. We choose Gamma priors $\alpha \sim \text{Ga}(b_\alpha, c_\alpha)$, $\beta \sim \text{Ga}(b_\beta, c_\beta)$ and $\tilde{m} \sim \text{Ga}(b_{\tilde{m}}, c_{\tilde{m}})$, since these parameters are restricted to $(0, \infty)$. Note that the probability density function of $U \sim \text{Ga}(b, c)$ is given by $\pi_U(u) = \big(c^b/\Gamma(b)\big)\, u^{b-1} e^{-cu}$ for $u > 0$ and $b, c > 0$. Additionally, we use Beta priors for the remaining parameters, which are restricted to [0,1], that is, $V \sim \text{Beta}(b, c)$, with probability density function given by, $\pi_V(v) = v^{b-1}(1-v)^{c-1}$. This family of distributions has the advantage of being the conjugate prior distribution to the Binomial likelihood (Altman, 1990). Using these prior specifications we have that:

$$\pi(\tilde{\boldsymbol{\theta}}) = \Big[\pi_\alpha\big(e^{\tilde{\alpha}}\big) \times e^{\tilde{\alpha}}\Big] \times \Big[\pi_\beta\big(e^{\tilde{\beta}}\big) \times e^{\tilde{\beta}}\Big] \times \pi_{\tilde{m}}(\tilde{m}) \times \pi_\nu(\nu) \times \pi_{\theta_R}(\theta_R) \times \pi_{\theta_F}(\theta_F).$$

We consider a hybrid Gibbs sampler to obtain samples from the marginal posterior distribution of each parameter. Some of the full conditionals are not given in closed form; for these we resort to HMC, as described in Section 1.2.3. The hidden carriage process is simulated from its full conditional using the forward filtering backwards sampling algorithm. Our prior specifications for ν, θ_R and θ_F are conjugate and therefore these parameters are directly sampled from their Beta full conditionals. The remaining model parameters $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{m}$ are jointly updated with HMC. An overview of the MCMC algorithm can be found in Algorithm 3. The main interest lies in the update of the hidden process $\mathbf{X}$, for which we give the details in the following Section 3.4.1; for the remaining updates the reader can refer to Appendix B.1.

Algorithm 3: MCMC algorithm for the hidden Markov model

```
Initialise $\tilde{\boldsymbol{\theta}}^{(0)} = \left(\tilde{\alpha}^{(0)}, \tilde{\beta}^{(0)}, \tilde{m}^{(0)}, \nu^{(0)}, \theta_R^{(0)}, \theta_F^{(0)}\right)$ and generate $\mathbf{X}^{(0)} \sim \pi\left(\mathbf{X} \mid \tilde{\boldsymbol{\theta}}^{(0)}\right)$;
for $j$ = 1, 2, ..., $J$ do
    Draw $\mathbf{X}^{(j)} \sim \pi\left(\mathbf{X} \mid \tilde{\alpha}^{(j-1)}, \tilde{\beta}^{(j-1)}, \tilde{m}^{(j-1)}, \nu^{(j-1)}, \theta_R^{(j-1)}, \theta_F^{(j-1)}, \mathbf{Y}\right)$;
    Perform a HMC to update $\left(\tilde{\alpha}^{(j)}, \tilde{\beta}^{(j)}, \tilde{m}^{(j)}\right)$ according to $\pi\left(\tilde{\alpha}, \tilde{\beta}, \tilde{m} \mid \mathbf{X}^{(j)}, \nu^{(j-1)}, \theta_R^{(j-1)}, \theta_F^{(j-1)}, \mathbf{Y}\right)$;
    Draw $\nu^{(j)} \sim \pi\left(\nu \mid \mathbf{X}^{(j)}, \tilde{\alpha}^{(j)}, \tilde{\beta}^{(j)}, \tilde{m}^{(j)}, \theta_R^{(j-1)}, \theta_F^{(j-1)}, \mathbf{Y}\right)$;
    Draw $\theta_R^{(j)} \sim \pi\left(\theta_R \mid \mathbf{X}^{(j)}, \tilde{\alpha}^{(j)}, \tilde{\beta}^{(j)}, \tilde{m}^{(j)}, \nu^{(j)}, \theta_F^{(j-1)}, \mathbf{Y}\right)$;
    Draw $\theta_F^{(j)} \sim \pi\left(\theta_F \mid \mathbf{X}^{(j)}, \tilde{\alpha}^{(j)}, \tilde{\beta}^{(j)}, \tilde{m}^{(j)}, \nu^{(j)}, \theta_R^{(j)}, \mathbf{Y}\right)$;
end
```

3.4.1 Updating the hidden states

Using the Markov property of the model, one can use the forward filtering backwards sampling (FFBS) method, as described by Carter and Kohn (1994), to perform single block sampling for each $\mathbf{X}^p$, $p = 1, 2, \ldots, P$ independently one of another. The method is based on the factorisation of $\mathbb{P}(\mathbf{X}^p \mid \mathbf{Y}^p, \tilde{\boldsymbol{\theta}})$, as:

$$\mathbb{P}(\mathbf{X}^p \mid \mathbf{Y}^p, \tilde{\boldsymbol{\theta}}) = \mathbb{P}\left(\mathbf{X}_{T^p}^{[1:n_c^p, p]} \mid \mathbf{Y}^p, \tilde{\boldsymbol{\theta}}\right) \prod_{t=1}^{T^p-1} \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} \mid \mathbf{X}_{t+1:T^p}^{[1:n_c^p, p]}, \mathbf{Y}^p, \tilde{\boldsymbol{\theta}}\right).$$

Under the conditional independence assumptions of our model (see Figure 3.2), the expression $\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} \mid \mathbf{X}_{t+1:T^p}^{[1:n_c^p, p]}, \mathbf{Y}^p, \tilde{\boldsymbol{\theta}}\right)$ reduces to $\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} \mid \mathbf{X}_{t+1}^{[1:n_c^p, p]}, \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \tilde{\boldsymbol{\theta}}\right)$, for $t = 1, 2, \ldots, (T^p - 1)$, and can be calculated:

$$\begin{aligned}
&\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k} \mid \mathbf{X}_{t+1}^{[1:n_c^p, p]} = \mathbf{x}_{t+1}^{[1:n_c^p, p]}, \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \tilde{\boldsymbol{\theta}}\right) \\
&\quad \propto \mathbb{P}\left(\mathbf{X}_{t+1}^{[1:n_c^p, p]} = \mathbf{x}_{t+1}^{[1:n_c^p, p]} \mid \mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k}, \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \tilde{\boldsymbol{\theta}}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k} \mid \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \tilde{\boldsymbol{\theta}}\right) \\
&\quad = \mathbb{P}\left(\mathbf{X}_{t+1}^{[1:n_c^p, p]} = \mathbf{x}_{t+1}^{[1:n_c^p, p]} \mid \mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k} \mid \mathbf{Y}^{[1:n_c^p, p]}\right), \quad (3.9)
\end{aligned}$$

where $\mathbf{k} \in \mathcal{X}_s^{n_c^p} = \{0, 1\}^{n_c^p}$. Therefore, the algorithm is based upon a forward recursion which calculates $\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} \mid \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \right)$ (the second mass function in Equation (3.9)), called the filtered probabilities, for $t = 1, 2, \ldots, T^p$, using the recursive filtering equations by Anderson and Moore (1979). This is followed by a backward simulation step that first generates $\mathbf{X}_{T^p}^{[1:n_c^p,p]}$ from $\mathbb{P}\left(\mathbf{X}_{T^p}^{[1:n_c^p,p]} \mid \mathbf{Y}^p, \tilde{\boldsymbol{\theta}}\right)$ and then simulates the remaining $\mathbf{X}_t^{[1:n_c^p,p]}$'s by progressing backwards, simulating in turn $\mathbf{X}_t^{[1:n_c^p,p]}$ from $\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,p]} \mid \mathbf{X}_{t+1}^{[1:n_c^p,p]}, \mathbf{Y}_{1:t}^{[1:n_c^p,p]}, \tilde{\boldsymbol{\theta}}\right)$, for $t = (T^p - 1), (T^p - 2), \ldots, 1$.

More specifically, the forward recursion is initialised at $t = 1$ with:

$$\mathbb{P}\left(\mathbf{X}_1^{[1:n_c^p,p]} = \mathbf{k} \mid \mathbf{Y}_1^{[1:n_c^p,p]}, \tilde{\boldsymbol{\theta}}\right) = \frac{\mathbb{P}\left(\mathbf{X}_1^{[1:n_c^p,p]} = \mathbf{k} \mid \nu\right) f_{\mathbf{k}}\left(\mathbf{y}_1^{[1:n_c^p,p]} \mid \boldsymbol{\vartheta}\right)}{\sum\limits_{\boldsymbol{\omega} \in \mathcal{X}_s^{n_c^p}} \mathbb{P}\left(\mathbf{X}_1^{[1:n_c^p,p]} = \boldsymbol{\omega} \mid \nu\right) f_{\boldsymbol{\omega}}\left(\mathbf{y}_1^{[1:n_c^p,p]} \mid \boldsymbol{\vartheta}\right)},$$

for $\mathbf{k} \in \mathcal{X}_s^{n_c^p}$. Then for $t = 2, 3, \ldots, T^p$ in a forward recursion, we first compute the one-step ahead predictive probabilities, by the law of total probability,

$$\begin{aligned} &\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,p]} = \mathbf{k} \mid \mathbf{Y}_{1:t-1}^{[1:n_c^p,p]}, \tilde{\boldsymbol{\theta}}\right) = \\ &\sum_{\boldsymbol{\omega} \in \mathcal{X}_s^{n_c^p}} \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,p]} = \mathbf{k} \mid \mathbf{X}_{t-1}^{[1:n_c^p,p]} = \boldsymbol{\omega}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_{t-1}^{[1:n_c^p,p]} = \boldsymbol{\omega} \mid \mathbf{Y}_{1:t-1}^{[1:n_c^p,p]}, \tilde{\boldsymbol{\theta}}\right), \end{aligned} \tag{3.10}$$

and then we compute the filtered probabilities,

$$\begin{aligned} &\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,p]} = \mathbf{k} \mid \mathbf{Y}_{1:t}^{[1:n_c^p,p]}, \tilde{\boldsymbol{\theta}}\right) \\ &= \frac{f_{\mathbf{k}}\left(\mathbf{y}_t^{[1:n_c^p,p]} \mid \boldsymbol{\vartheta}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,p]} = \mathbf{k} \mid \mathbf{Y}_{1:t-1}^{[1:n_c^p,p]}, \tilde{\boldsymbol{\theta}}\right)}{\sum\limits_{\boldsymbol{\omega} \in \mathcal{X}_s^{n_c^p}} f_{\boldsymbol{\omega}}\left(\mathbf{y}_t^{[1:n_c^p,p]} \mid \boldsymbol{\vartheta}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,p]} = \boldsymbol{\omega} \mid \mathbf{Y}_{1:t-1}^{[1:n_c^p,p]}, \tilde{\boldsymbol{\theta}}\right)}, \quad \text{for } \mathbf{k} \in \mathcal{X}_s^{n_c^p}. \end{aligned} \tag{3.11}$$

Once the filtered probabilities have been calculated and stored in a forward sweep, the hidden states can be simulated in a backward sweep starting with the simulation of a value for $\mathbf{X}_{T^p}^{[1:n_c^p,p]}$ from the filtered probability at time T^p, $\mathbb{P}\left(\mathbf{X}_{T^p}^{[1:n_c^p,p]} \mid \mathbf{Y}^p, \tilde{\boldsymbol{\theta}}\right)$. Then for $t = (T^p - 1), (T^p - 2), \ldots, 1$ we iteratively sample a value for $\mathbf{X}_t^{[1:n_c^p,p]}$ given our simulated value for $\mathbf{X}_{t+1}^{[1:n_c^p,p]}$, from:

$$\begin{aligned} &\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,p]} = \mathbf{k} \mid \mathbf{X}_{t+1}^{[1:n_c^p,p]} = \mathbf{x}_{t+1}^{[1:n_c^p,p]}, \mathbf{Y}_{1:t}^{[1:n_c^p,p]}, \tilde{\boldsymbol{\theta}}\right) \\ &\quad = \frac{\mathbb{P}\left(\mathbf{X}_{t+1}^{[1:n_c^p,p]} = \mathbf{x}_{t+1}^{[1:n_c^p,p]} \mid \mathbf{X}_t^{[1:n_c^p,p]} = \mathbf{k}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,p]} = \mathbf{k} \mid \mathbf{Y}_{1:t}^{[1:n_c^p,p]}, \tilde{\boldsymbol{\theta}}\right)}{\sum\limits_{\boldsymbol{\omega} \in \mathcal{X}_s^{n_c^p}} \mathbb{P}\left(\mathbf{X}_{t+1}^{[1:n_c^p,p]} = \mathbf{x}_{t+1}^{[1:n_c^p,p]} \mid \mathbf{X}_t^{[1:n_c^p,p]} = \boldsymbol{\omega}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,p]} = \boldsymbol{\omega} \mid \mathbf{Y}_{1:t}^{[1:n_c^p,p]}, \tilde{\boldsymbol{\theta}}\right)}, \end{aligned}$$

for $\mathbf{k} \in \mathcal{X}_s^{n_c^p}$.

3.4.2 Implementation details

In what follows, we use the following prior specifications unless otherwise stated. We assume mutually independent prior distributions for model parameters, specifically that α, $\beta \sim \text{Ga}(1, 1)$, $\tilde{m} \sim \text{Ga}(0.01, 0.01)$ and ν, θ_R, $\theta_F \sim \text{Beta}(1, 1)$, which is the same as the Uniform distribution on the interval [0,1]. These priors are chosen in order to be weakly informative, that is, provide limited information about the values of the unknown parameters.

In all implementations we run the MCMC for a total of 75,000 iterations. We then discard the first 25,000 as a burn-in period and of the remaining 50,000 we record the output at every 10 iterations to obtain a samples of size 5,000 from the posterior distribution. The convergence of the chains is assessed by visual inspection of the posterior traceplots and also tested with the Geweke criteria (Geweke, 1992).

3.5 Simulation studies

A simulation study is conducted to evaluate the performance of the proposed estimation method for the parameters of the model. To study the impact of disease characteristics on algorithm performance, we specify 3 different scenarios for the spread of infection through a pen. In all 3 scenarios, the external and within-pen colonisation rates, as well as the initial probability of colonisation remained constant. More specifically, simulated data are generated with external colonisation rate α equal to 0.01, within-pen transmission rate β equal to 0.0125 and the initial probability of infection is set to $\nu = 0.1$. For the mean duration of carriage, the following three scenarios are considered: $m = 5$ (S1), $m = 10$ (S2) and $m = 15$ (S3), which correspond to low, moderate and long disease duration. Also, the RAMS and faecal test sensitivities are assumed to be 0.8 and 0.5, respectively, same for all different scenarios.

The simulation proceeds as follows. We generate a complete dataset of carriage status for 127 days, one for each scenario, which corresponds to the average observation period of two motivating datasets described in Section 2.2. Given the true carriage states, RAMS and faecal test results are simulated for each individual for the entire period of 127 days. For each scenario, we then extract two separate longitudinal datasets from the complete set of test observations, with sampling frequency twice per week and once every two weeks, resulting in 37 and 10 pairs of RAMS and faecal samples for each individual respectively. The intervals resemble the actual sampling frequencies in the two motivating longitudinal studies. Using 2 different sampling intervals allows us to investigate how testing frequency can alter inferences, especially since the complete disease data are shared. To avoid sampling biases, for each scenario and sampling interval, 40 epidemics were simulated in a subset of 20 pens, each containing 8 individuals.

In Figure 3.3 we show the posterior medians along with 95% credible intervals over the 40 realisations of the two sampling schemes, under the 3 different scenarios. The true parameter values are also shown in the figure. Overall, we see that all credible intervals include the true parameter values suggesting that our algorithm succeeds at estimating these parameters. As we expected, the performance depends on the sampling frequency. In particular, longer sampling intervals (in green) are associated with greater variability. This is due to fact that there are less observations when the sampling interval is sparse and hence less information to impute the hidden states.

When samples are taken twice per week, the true mean duration of the disease does not greatly affect our parameter estimates. However, issues arise when the sampling interval is much longer compared to the mean duration of the disease. This can be seen in scenario 1, when the mean duration is 5 days and the sampling interval is set to 2 weeks. In this case, the duration of colonisation is underestimated and both the external and within-pen colonisation rates are overestimated. The reason for this behaviour is demonstrated in Figure 3.4. In this figure we plot the posterior probability of colonisation for all individuals in a randomly selected pen; those are obtained for every day as the proportion of iterations of the MCMC in which the carriage state of a given individual was imputed as 1 (colonised) by the FFBS algorithm. We see that there are occasions where an individual is colonised after a testing day and is subsequently cleared before the next samples are taken. Therefore, small periods of carriage remain completely unobserved. As an example, see individual 5 at the period between days 43 and 57. In contrast, such incidents are less prevalent with a sampling frequency of twice per week, see Figure 3.5.

Convergence of the hidden carriage states is assessed by visual inspection of posterior colonisation probabilities over the sampling period, shown for example in Figure 3.5 for one randomly selected pen. For reference, we also display the observed results of RAMS (red) and faecal (blue) tests taken twice per week, as well as the daily true carriage states (black squares). The estimated probabilities match the true states very well, indicating that the algorithm accurately reproduces the dynamics of the disease. Notice that when an individual has a positive RAMS and/or faecal sample then the posterior probability of colonisation is equal to 1 due to the assumption of perfect test specificity: a subject whose test indicated as positive actually has the disease. On the other hand, when both test results are negative

Figure 3.3: Comparison of the distributions of the posterior median estimates of parameters based on 40 simulated datasets with different sampling intervals under three epidemic scenarios. For each scenario the red dashed lines indicate the true values of the corresponding model parameter. Boxplots give the quantiles 2.5%, 25%, 50%, 75%, and 97.5%, respectively.

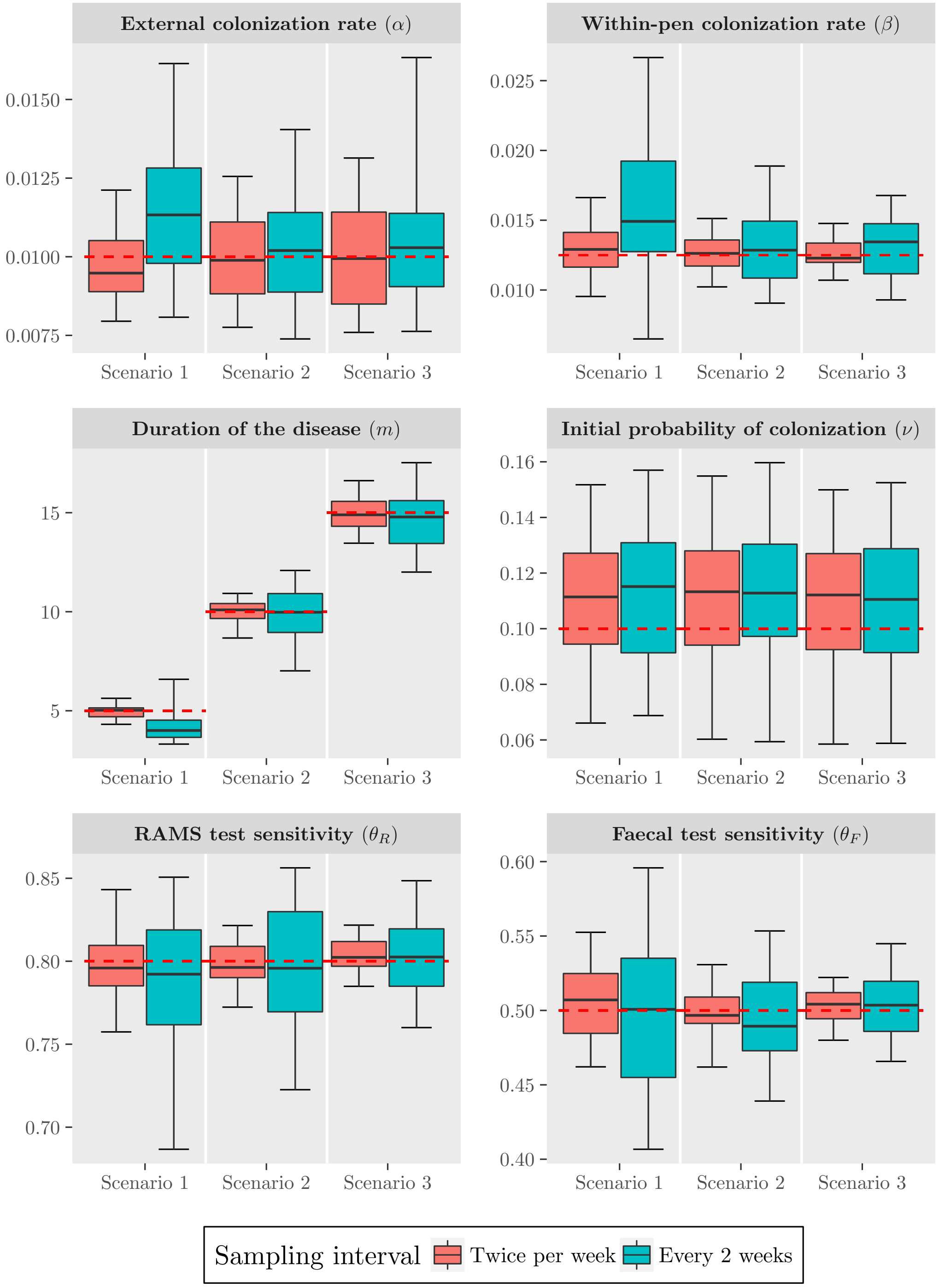


Figure 3.4: Posterior probability of colonisation (grey solid line) for individuals in the simulated dataset under the first scenario ($m = 5$), over the entire sampling period of 127 days. Black squares represent the true colonisation states (1 for colonised, 0 for non-carrier). For reference we also show test results taken every two weeks; "·" indicates negative sample and "+" indicates that the sample was positive. White vertical lines represent the days in which samples were taken.

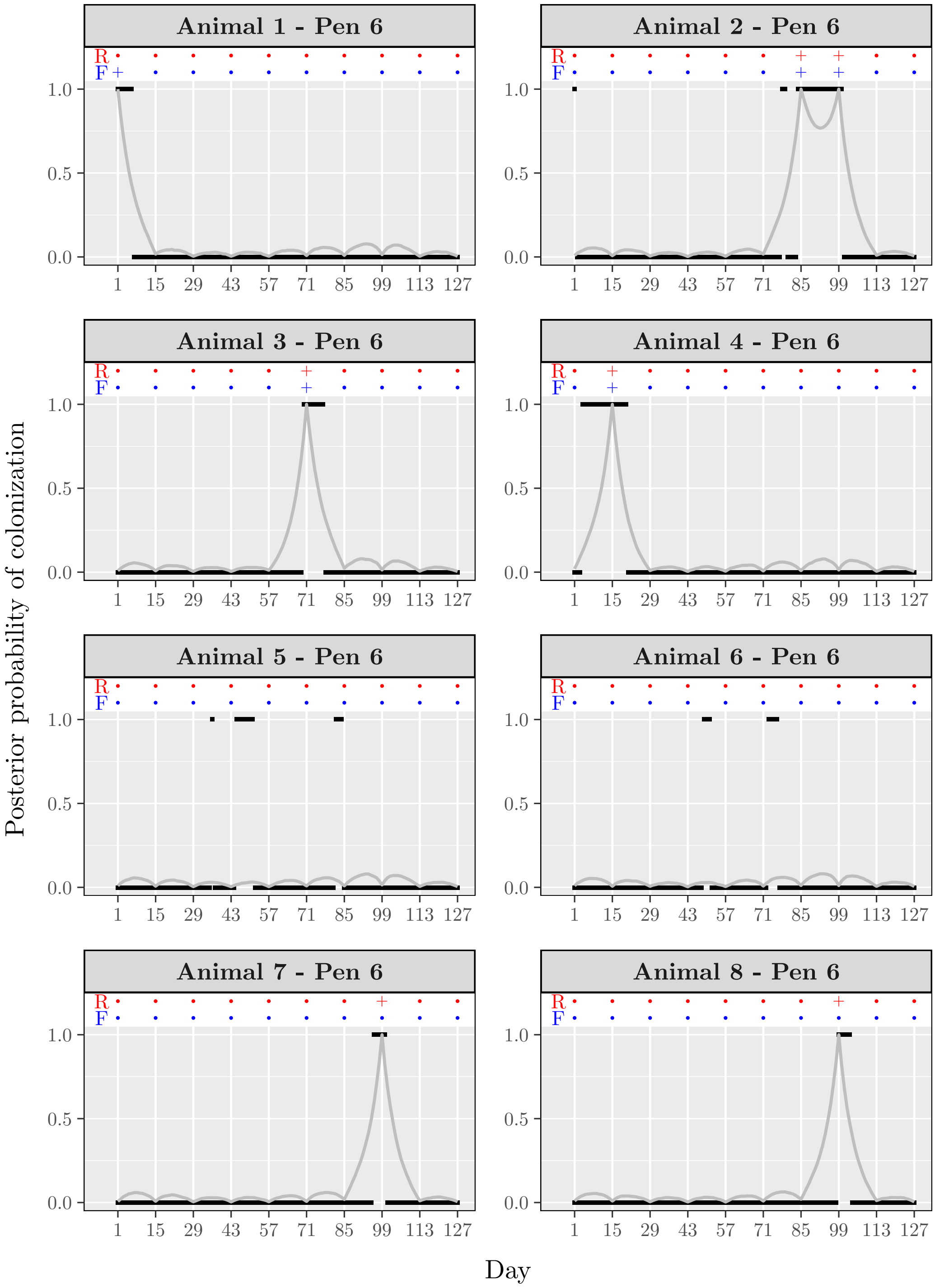
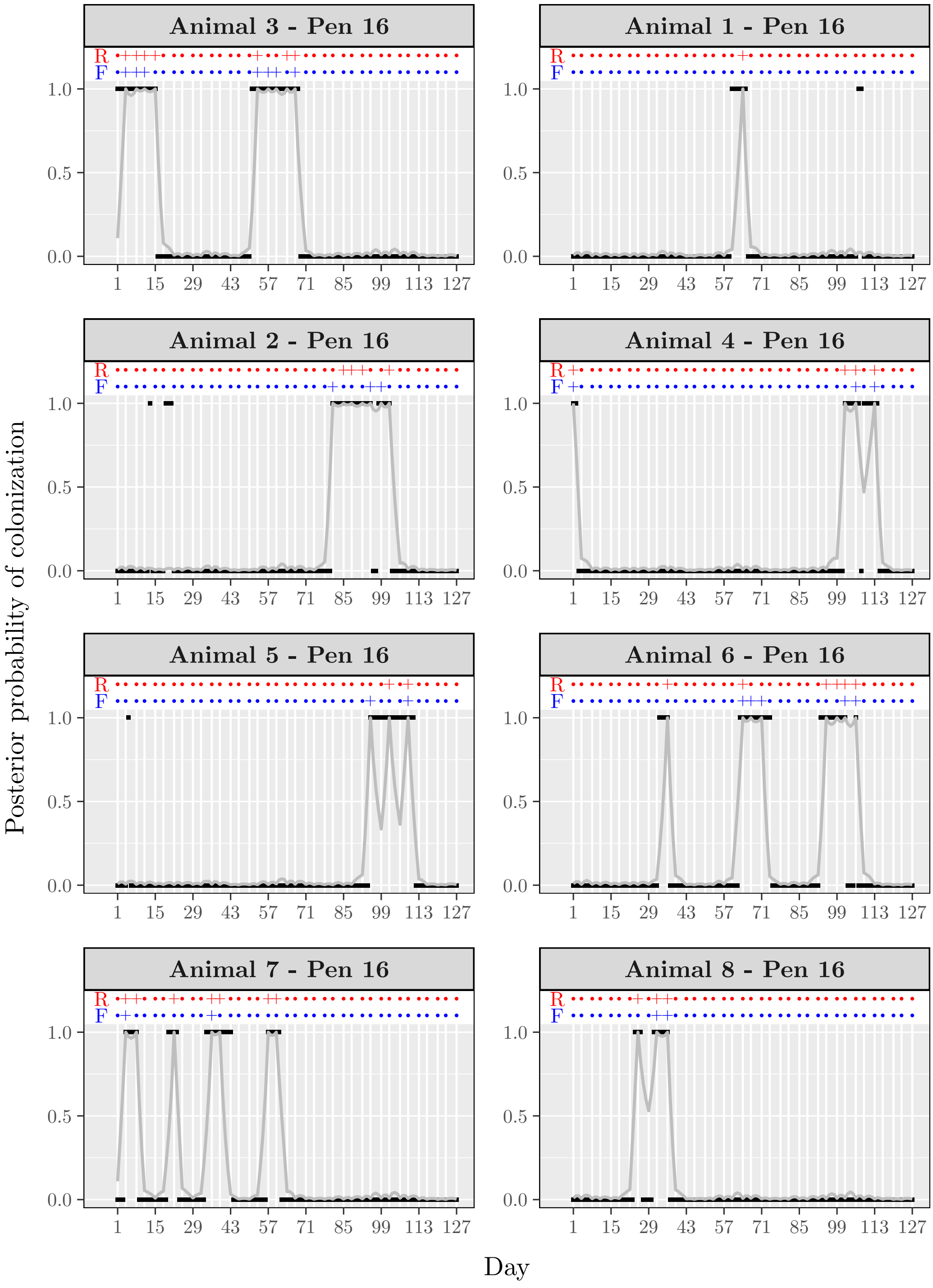


Figure 3.5: Posterior probability of colonisation (grey solid line) for individuals in the simulated dataset under the first scenario ($m = 5$), over the entire sampling period of 127 days. Black squares represent the true colonisation states (1 for colonised, 0 for non-carrier). For reference we also show test results taken twice per week; "·" indicates negative sample and "+" indicates that the sample was positive. White vertical lines represent the days in which samples were taken.



the posterior probability of infection can be any value between and including 0 and 1. In addition, sequences of positive results separated by negative results can either attributed to lack of sensitivity or to true conversions. This results in our method being uncertain as to whether an individual is an undetected carrier or has re-acquired the disease and hence we observe spikes in the posterior probabilities, e.g. animal 5 at day 99 and animal 8 at day 29, respectively.

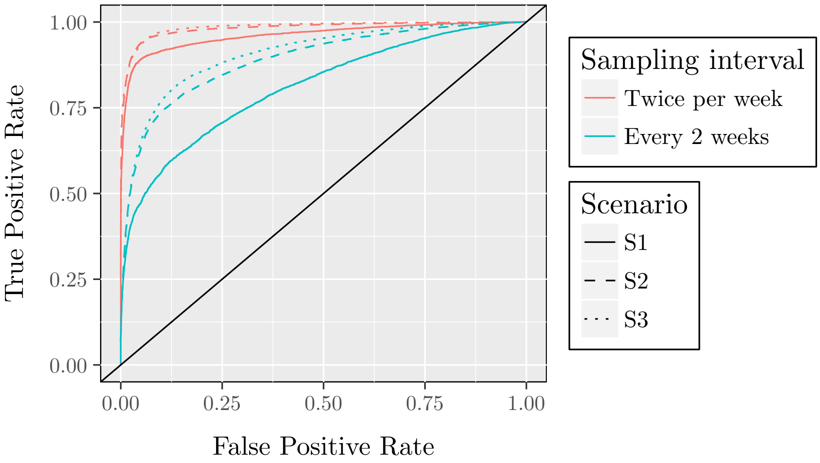
Given the posterior probability of colonisation for an individual at any time during the study period, one can classify the individual as carrier if the probability exceeds a given threshold. A colonised individual correctly classified by the algorithm as carrier is a true positive finding whereas a false positive occurs when the method misclassifies a susceptible individual as a carrier. Therefore we can further evaluate the performance of our method for imputing the hidden carriage states using the receiver operating characteristics (ROC) curve. The ROC curve is a plot of the true positive rates against false positive rates, calculated over different threshold values on the posterior colonisation probabilities. We consider the area under curve (AUC) as a measure of performance of our method. In Figure 3.6 we plot the median ROC curves produced over 40 replicates of each scenario and sampling interval; the corresponding AUCs are presented in Table 3.1 along with 95% credible intervals. With the twice per week sampling frequency, the AUC exceeds 96%, confirming the good performance of the method that was observed in Figure 3.5. However, with longer sampling interval (once every 2 weeks) the performance drops. The effect of m is clearly demonstrated in Table 3.1: increasing m associates with higher AUC. This trend is stronger with sparse sampling frequency.

Further simulation studies were constructed in order to resemble our real data. The simulations were designed to have the following attributes the same as the data: total number of pens, total number of individuals per pen, sampling days and missing observations per individual. The results are similar to the results obtained in the simulation study shown earlier and can be found in Figure B.1 of

Table 3.1: Median area under the ROC curve, as obtained from 40 simulated datasets under different sampling intervals and 3 duration scenarios S1-S3. The 95% credible intervals are shown in parentheses.

Scenario	**Sampling interval**	
	Twice per week	**Every 2 weeks**
S1	0.96 (0.95, 0.97)	0.81 (0.75, 0.84)
S2	0.98 (0.97, 0.99)	0.89 (0.87, 0.91)
S3	0.98 (0.98, 0.99)	0.91 (0.89, 0.93)

FIGURE 3.6: Median ROC curves obtained from 40 simulated datasets, under different sampling intervals and scenarios. Red lines represent a sampling interval of twice per week whereas green lines correspond to one testing day per 2 weeks. Scenarios 1, 2 and 3 are represented by solid, dashed and dotted lines respectively.



the Appendix.

3.6 Applications

In this section we apply the Bayesian approach presented in Section 3.4 to the real data analysis problems. The datasets that we consider are the *E. coli* O157:H7 datasets 1 and 2 described in Section 2.2. This allows us to obtain estimates for epidemiologically important parameters and draw conclusions of the prevalence of *E. coli* O157:H7 in cattle. Except from inferential results our study includes sensitivity analyses, as well as model checking which is based on comparing predictive distribution with the observed data.

Prior specifications and details of the MCMC implementation are given in Section 3.4.2, except for when stated otherwise. A point which is worth emphasising is that cattle withdrawn from the study are removed from the model on their dates of drop-out. Therefore missing carriage states from these animals are not imputed since they do not play further role in the spread of the epidemic (see Section B.2 of the Appendix for more details).

3.6.1 Dataset 1

3.6.1.1 Results

Convergence of the chains is done by visual inspection of traceplots for the model parameters, shown in Figure B.2 of the Appendix. These plots indicate that the model behaves well with good mixing for all 6 parameters, which appear to reach stationarity. We check that estimates are robust to a change in the initial values by running three different Markov chains; all runs result to the same posterior densities as shown in Figures B.4 and B.3 of the Appendix.

Posterior summaries and parameter interpretations are given in Table 3.2. The posterior median of the colonisation rate between an infected and a susceptible animal within the same pen in one day is 0.011 (95% CI, 0.007 to 0.014). This implies that a susceptible animal acquires infection from a given infected animal on average every 92 days (1/0.011). The posterior median rate of external transmission is 0.009 per day (95% CI, 0.006 to 0.012); a susceptible individual acquires infection from other environmental sources on average every 111 days (1/0.009). We find that the ratio $\beta/\alpha = 1.2$. Since the value is close to 1 we conclude that infectious pressure exercised by one single carrier of *E. coli* O157:H7 within a pen is as strong as all of the external sources of colonisation. The mean duration of colonisation is 9.4 days (95% CI, 8 to 11). The initial probability of colonisation is 0.099 with a 95% credible interval of 0.056 to 0.154.

Regarding the diagnostic test sensitivities, we estimate that RAMS test has probability of successful disease detection equal to 77.7% (95% CI, 73.1 to 81.6%) while the sensitivity of the faecal test is estimated to be 46.5% (95% CI, 42.1 to 50.7%), indicating that the latter leads to many false negatives. These results suggest that the RAMS technique achieves a much higher sensitivity of detecting *E. coli* O157:H7 in cattle than the faecal method. This finding is consistent with previous studies (Greenquist *et al.*, 2005; Rice *et al.*, 2003). Overall, our parameter estimates are in close agreement with results obtained by Spencer *et al.* (2015) who previously analysed the same data.

Of particular interest is the estimation of the basic reproduction number R_0 of *E. coli* O157:H7, which is defined as the expected number of secondary infections produced from one infected individual when introduced into a susceptible population (Anderson and May, 1991). In finite populations, dealing with R_0 presents a number of difficulties. For each pen p, we define R_0 as $m\left(1-e^{-\beta}\right)\left(n_c^p-1\right)$, where m is the expected lifetime of infection, $\left(1-e^{-\beta}\right)$ is the probability of transmission given contact of a susceptible and an infected individual and $n_c^p - 1$ is the number of

Table 3.2: Interpretation and posterior medians (posterior standard deviations) of the model parameters for the analysis of *E. coli* O157:H7 dataset 1. Values in brackets indicate 95% credible interval (CI).

Interpretation	**Symbol**	**Posterior median (sd)** **[95% CI]**
External colonisation rate (days^{-1})	α	0.009 (0.001) [0.006, 0.012]
Within-pen colonisation rate (days^{-1})	β	0.011 (0.002) [0.007, 0.014]
Mean colonisation duration (days)	m	9.419 (0.757) [8.032, 11.04]
Probability colonised initially	ν	0.099 (0.025) [0.056, 0.154]
Sensitivity of RAMS	θ_R	0.777 (0.022) [0.731, 0.816]
Sensitivity of faecal test	θ_F	0.465 (0.022) [0.421, 0.507]

susceptible individuals in the pen, excluding the initially infected. A weakness of the definition that we choose is that if there are multiple infectious contacts after the first one, the contacted individual may still be infected and therefore this is the maximum value that basic reproduction number can take. However, this is a relatively rare event and therefore we expect that the estimated R_0 will be only slightly higher compared to estimates obtained using other definitions of the basic reproduction number. A key application of R_0 is its use as a threshold parameter, such that a major outbreak can only occur if R_0 is more than 1. In our case, the within pen basic reproduction ratio is estimated 0.69 (95% CI, 0.50 to 0.91) which is below the threshold at which new infections tend to increase and thus a major outbreak can not occur within the pen. This implies that *E. coli* O157:H7 is not able to successfully spread in a cattle pen without reintroduction from sources external to the pen.

Since our approach involves estimating unobserved infection status with FFBS algorithm, we can also estimate quantities such as the prevalence of the disease. The overall prevalence of *E. coli* O157:H7 for our data, which takes into account colonised animals that failed to test positive, is estimated to be 14.57% (95% CI, 13.99 to 15.22%). Relying on the available results for the RAMS and faecal tests, we further create the plot of the posterior probability of colonisation for every animal in every day of the study. As an example, in Appendix B.4 we show these quantities for pens 3 and 19 in Figures B.5 and B.6 respectively.

3.6.1.2 Model Checking

Following Gelman *et al.* (1996), we use posterior predictive checks to assess our model fit. This is done by comparing posterior predictive data to the actual observations. The diagnostics that we consider are presented as follows. The first is the detection duration using a particular diagnostic test, which is defined as the number of consecutive days an animal is tested positive by the test. Since visits were not conducted every day, the duration is calculated by determining the interval (in days) between the first and last consecutive sample visits that yielded positive samples and adding 1 day. The second and third diagnostics are the total numbers of positive test results and the total number of animals that never tested as positive for a particular test. The latter diagnostic is interesting from an epidemic perspective since it can be used to assess a potential weakness of our model, namely that all individuals are homogeneous in susceptibility.

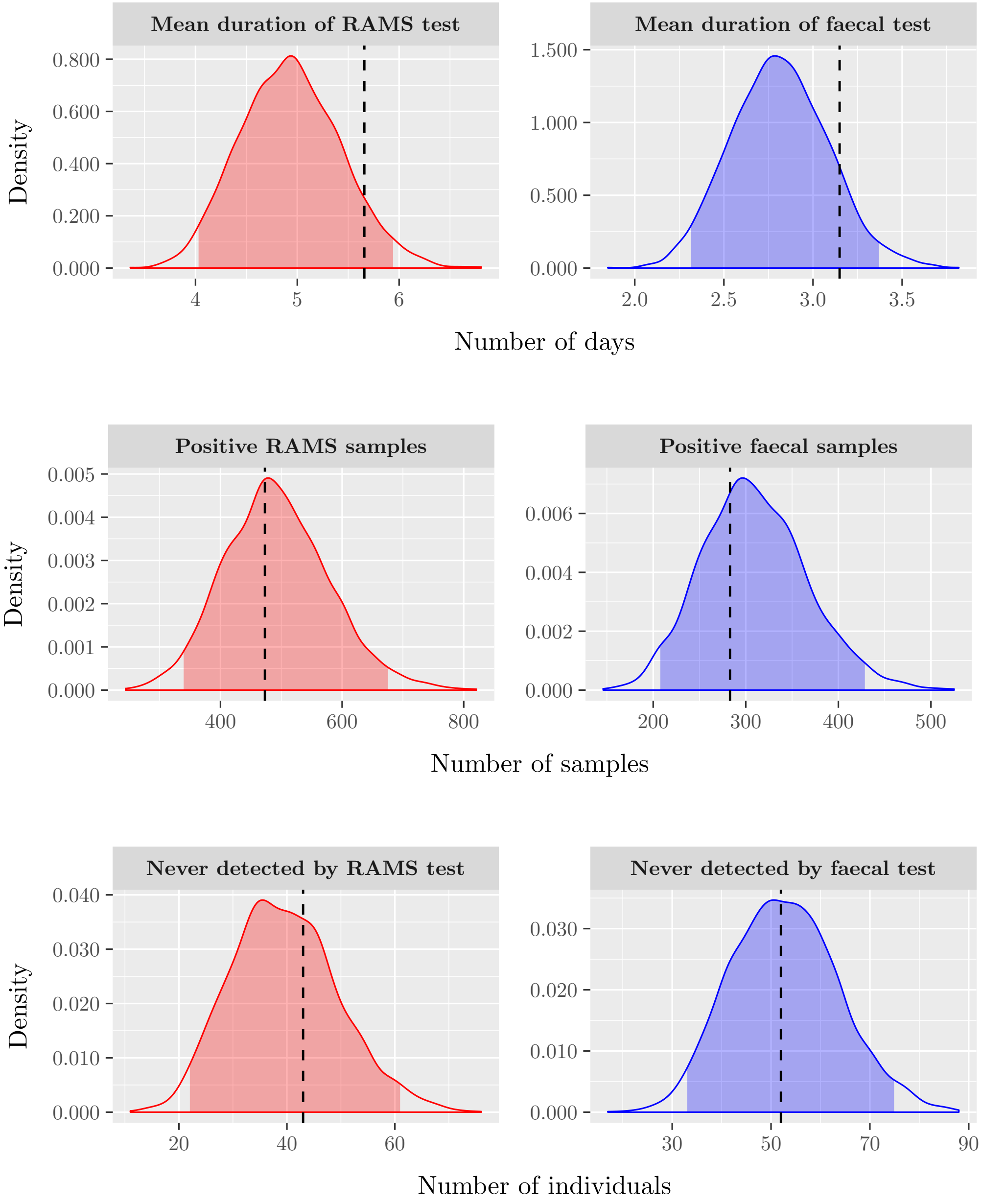
For each iteration of the MCMC algorithm, we simulate faecal and RAMS samples from individual cattle according to the actual sampling dates and conditional on the model parameter samples at this iteration and calculate the quantities described above. This is done for 5,000 iterations in total. We then plot distributions of these quantities and check whether the observed data values are placed within the 95% posterior predictive intervals.

Results for the three diagnostics can be found in Figure 3.7. For the RAMS test, the data mean of detection duration is 5.66, which falls inside the 95% CI of the predictive distribution (4.02 to 5.94). Similarly, for the faecal test the mean detection duration of the real data is 3.15 and is again placed within the 95% CI of the posterior predictive samples (2.31 to 3.37). However, both of these quantities lie close to the upper limit border, which might suggest that the Geometric distribution is inappropriate to model the duration of colonisation. We might need to consider a more flexible distribution for the colonisation period, for example the Negative Binomial distribution, which is an extension of the Geometric model. The remaining observed data values are all placed well within the 95% credible intervals in the predictive distribution, and thus we conclude that there is no evidence of heterogeneities in susceptibility between individuals. Overall, the model fit appears adequate.

3.6.1.3 Prior sensitivity

We repeat the analysis of dataset 1 with different prior specifications to investigate the effects on posterior estimates. Sensitivity is only examined for three parameters

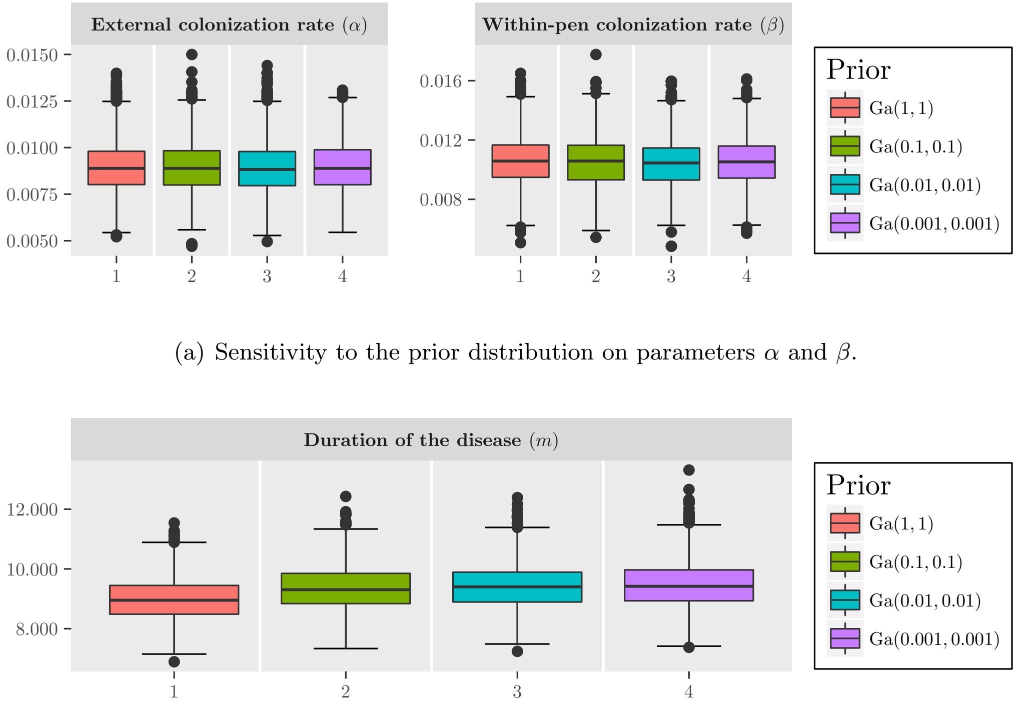
Figure 3.7: Model assessment plots for *E. coli* O157:H7 dataset 1. Posterior predictive distribution of the mean duration (top panel) and total number (middle panel) of positive samples, and total number of animals that never tested as positive (bottom panel) for RAMS (left panel) and faecal samples (right panel), respectively. Black dashed line indicate the observed value of the corresponding summary. Shaded area corresponds to the 95% credible interval. The results are based on 5,000 posterior predictive simulations having the same structure as in the original dataset.



namely the mean duration of colonisation, external and within-pen transmission rates. For each of these parameters, we use 4 different Gamma priors, 1-4, with constant mean equal to 1 and variance equal to 1, 10, 100 and 1000 respectively. Each time, the prior of only one parameter is changed.

In Figure 3.8 we plot the marginal posterior distributions of the parameters under investigation, for the different choice of priors. No major change is observed in the posterior median and quantiles of neither external nor within-pen transmission rates. The duration of the colonisation is somewhat sensitive to its prior distribution: the posterior distribution is shifted roughly 0.5 days upwards when replacing the most informative Ga(1, 1) with one of the other 3 alternatives.

Figure 3.8: Sensitivity to the prior distribution on the colonisation rates and the mean colonisation duration for *E. coli* O157:H7 dataset 1. We use 4 different Gamma priors, 1-4, with constant mean equal to 1 and variance equal to 1, 10, 100 and 1000 respectively. Each time, the prior of only one parameter is changed.



(a) Sensitivity to the prior distribution on parameters α and β.

(b) Sensitivity to the prior distribution on parameter m.

3.6.2 Dataset 2

3.6.2.1 Implementation details

As already mentioned, the results for the two diagnostic tests for this study are not fully available; we instead observe the $Y_t^{[c,p]} = \max\left(R_t^{[c,p]}, F_t^{[c,p]}\right)$ for each individual c in pen p on a given observation day $t \in O^{c,p}$. Therefore, the conditional distribution of the observed data given the hidden carriage states is given in terms of the sensitivities of both diagnostic tests as follows:

$$f_{x_t^{[c,p]}}\left(y_t^{[c,p]} \mid \theta_R, \theta_F\right) = \begin{cases} \Big(1-(1-\theta_F)(1-\theta_R)\Big) & \text{if } x_t^{[c,p]} = 1 \text{ and } y_t^{[c,p]} = 1, \\ (1-\theta_F)(1-\theta_R) & \text{if } x_t^{[c,p]} = 1 \text{ and } y_t^{[c,p]} = 0, \\ 0 & \text{if } x_t^{[c,p]} = 0 \text{ and } y_t^{[c,p]} = 1, \\ 1 & \text{if } x_t^{[c,p]} = 0 \text{ and } y_t^{[c,p]} = 0. \end{cases}$$

The test sensitivities are no longer identifiable so we fix these parameters θ_R and θ_F to 0.729 and 0.686, respectively, following the initial analysis of this study (including data that was not available to us) by Cernicchiaro *et al.* (2010).

3.6.2.2 Results

We run 3 chains starting from diverse initial values. Convergence is assessed via visual inspection of the sample chains (traceplots shown in Figure B.7 of the Appendix) which suggest that the chains have reached their common stationary distribution.

Marginal posterior density means, posterior standard deviations and 95% credible intervals for the 4 model parameters are shown in Table 3.3. The posterior median for the within-pen colonisation rate is 0.011 per day (95% CI, 0.008 to 0.014) which is identical to our estimate for the first dataset. Differences are found with respect to the remaining parameters, though. The estimated external transmission rate for this dataset is 0.004, roughly 2 times smaller compared to dataset 1. The ratio β/α is 2.75 which implies that the disease is more transmissible from within-pen contact. Mean duration of colonisation is estimated 6 days longer than dataset 1 (15.7 versus 9.4); however, the estimate is associated with larger variability. We find that 5% of the animals are colonised at the start of the study, half of the value found in dataset 1.

The within-pen basic reproduction number R_0 for this dataset is 1.03 with 95% credible interval of [0.76, 1.28]. The result suggests that the infection will be able to spread and be maintained in the pen in contrast to the first dataset. Finally, the prevalence of the disease is found 14.88% (95% CI $[14.01, 15.85]$), similar to dataset 1.

Table 3.3: Interpretation and posterior medians (posterior standard deviations) of the model parameters for the analysis of *E. coli* O157:H7 dataset 2. Values in brackets indicate 95% credible interval (CI).

Interpretation	**Symbol**	**Posterior median (sd) [95% CI]**
External colonisation rate (days^{-1})	α	0.004 (0.001) [0.003, 0.006]
Within-pen colonisation rate (days^{-1})	β	0.011 (0.002) [0.008, 0.014]
Mean colonisation duration (days)	m	15.69 (1.791) [12.59, 19.62]
Probability colonised initially	ν	0.049 (0.018) [0.022, 0.090]

3.6.2.3 Model Checking

In order to check the model adequation to the data, we use the same diagnostics as with dataset 1. In the simulations the observed test results are generated conditional on the simulated hidden states given the fixed test sensitivities, and the maximum value is recorded as done in the real data. Results provide no evidence against the fit of our model since all of the observed summary statistics lie within the 95% predictive credible intervals. Plots can be found in Figure B.8 of Appendix.

3.6.2.4 Prior sensitivity

The analysis of the dataset is repeated using 4 different values of the prior parameters as was done with the first dataset, and the results are shown in Figure B.9 of Appendix. No profound effect can be seen in the posterior distributions except for the duration of the colonisation (m) when the prior was a Ga$(1, 1)$. In particular, we find that the period of colonisation was 2 days shorter compared to the other priors. Overall, the results are consistent with ones obtained for dataset 1.

3.7 Discussion

In this chapter we have described a discrete-time non-homogeneous hidden Markov model that can be used to analyse the infection dynamics in longitudinal studies of *E. coli* O157:H7 among a population of cattle which is partitioned into pens. Although the application presented here considers *E. coli* O157:H7, it could be easily be adapted to model the transmission of other pathogens in other livestocks, or humans organised in households within a community. The hidden Markov model basis upon which our model is built provides a way to allow for the possibility of misclassification due to imperfect testing procedures. A significant merit of our approach is that it can easily handle missing data due to, for example, sparse sampling intervals or individual dropouts. Since the likelihood of the observed data given the model parameters is computationally intractable, we used a Bayesian method, in which the unobserved carriage process is included as an additional parameter, and inference is achieved with a hybrid Gibbs MCMC algorithm.

The model has been fitted to synthetic data and was found successful in accurately estimating the model parameters but also retrieve the hidden carriage process. A sensitivity analysis suggests that the accuracy of the estimates obtained by our method are less precise as the sampling interval becomes more sparse. Application on two longitudinal studies of *E. coli* O157:H7 in cattle has allowed us to estimate useful parameters regarding the dynamics of the disease. Also, it provided us with estimates of the sensitivity of test used in the studies. The results suggest that RAMS sampling is more sensitive at detecting *E. coli* O157:H7 in feedlot cattle than faecal sampling. Our finding is in agreement with the previous report by Rice *et al.* (2003).

There are several ways in which our model can be extended. First of all, our assumption of a Geometrically distributed colonisation period is not epidemiologically motivated for many diseases, although it makes the statistical analysis easier. A more flexible probability distribution is the Negative Binomial distribution, which allows the period for which an individual remains colonised to change over time. Moreover, in the above modelling framework we assumed that the risk of acquiring the infection from sources within and outside the pen is the same for all pens. However, in reality these parameters may vary between pens. An extension of this work will consider a hierarchical model in which each pen is allowed to have its own individual internal and external transmission parameters. Finally, one may consider an alternative model in which it is possible to have interactions between pens; e.g. dataset 2 where the pens share some border or their water supply. In the following chapters, we develop a set of stochastic epidemiological models that consider several of the aforementioned extensions and build a framework that enables comparison between these candidate models.

Chapter 4

Efficient Model Comparison Techniques For Models Requiring Large Scale Data Augmentation

This chapter is adapted from a joint paper with
Naif Alzahrani, Peter Neal, Simon E.F. Spencer and
Trevelyan J. McKinley (Touloupou *et al.*, 2015).

4.1 Introduction

The central pillar of Bayesian statistics is Bayes' Theorem. That is, given a parametric model $\mathcal{M}$ with parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$ and data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, the joint distribution of $(\boldsymbol{\theta}, \mathbf{y})$ satisfies

$$\pi(\boldsymbol{\theta} \mid \mathbf{y})\pi(\mathbf{y}) = \pi(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{4.1}$$

The four terms in Equation (4.1) are the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, the marginal likelihood or evidence $\pi(\mathbf{y})$, the likelihood $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ and the prior distribution $\pi(\boldsymbol{\theta})$. The terms on the right hand side of Equation (4.1) are usually easier to derive than those on the left hand side. The statistician has considerable control over the prior distribution and this can be chosen pragmatically to reflect prior beliefs and to be mathematically tractable. For many statistical problems the likelihood can easily be derived. However, the quantity of primary interest is usually the posterior distribution. Rearranging Equation (4.1) it is straightforward to obtain an expression for $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ so long as the marginal likelihood can be computed. This involves computing:

$$\pi(\mathbf{y}) = \int_{\boldsymbol{\theta}} \pi(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{4.2}$$

which is only possible analytically for a relatively small set of simple models.

A key solution to being unable to obtain an analytical expression for the posterior distribution is to obtain samples from the posterior distribution using MCMC (Metropolis *et al.*, 1953; Hastings, 1970). A major strength of MCMC is that it circumvents the need to compute $\pi(\mathbf{y})$ and this has led to its widespread use in Bayesian statistics over the last 25 years or so. However, Bayesian model choice typically requires the computation of Bayes Factors (Kass and Raftery, 1995) or posterior model probabilities, which are both functions of the marginal likelihoods for the competing models. In Chib (1995) a simple rewriting of Equation (4.1) was exploited to obtain estimates of the marginal likelihood using output from a Gibbs sampler. This has been extended in Chib and Jeliazkov (2001) and Chen (2005) to be used with the general Metropolis-Hastings algorithm. Importance sampling approaches to estimating the marginal likelihood have also been suggested (Gelfand and Dey, 1994), along with generalisations such as bridge sampling (Meng and Wong, 1996), which 'bridges' information from posterior and importance samples. More recently approaches have exploited the 'thermodynamic integral' such as power posterior methods Friel and Pettitt (2008). Alternative methods such as Sequential Monte Carlo (e.g. Zhou *et al.*, 2016) and nested sampling (Skilling, 2004) do not require any MCMC: computation of the marginal likelihood and samples from the posterior distribution are produced simultaneously. A potential drawback for many of the above approaches to marginal likelihood estimation is that it may not be obvious how to apply them efficiently to models incorporating large amounts of missing data.

It should be noted that there are model comparison techniques such as reversible jump MCMC (Green, 1995) which can be used to compare models without the need to compute the marginal likelihood. RJMCMC works well for nested models where it is straightforward to define a good transition rule for models with different parameters. However, in the case where we have large amounts of missing data it is often necessary to use some form of data augmentation technique, where the missing information is inferred alongside the other parameters of the model. Using RJMCMC becomes much harder in these cases since the dimension of the parameter space (including the augmented data) becomes large. This is exacerbated further when the missing information between the competing models has a different structure. In this latter case the use of intermediary (bridging) models (Karagiannis and Andrieu, 2013) to move between the models of interest is a possibility.

Here, we consider the problem of comparing a set of candidate models in a formal Bayesian model selection framework with focus on epidemiological data. In the context of epidemics, each model reflects an epidemiologically important hypothesis and the evidence in favour of each hypothesis (model) can be then measured using Bayes factors. However, as we explain in Chapter 3, data emerging from infectious disease outbreaks typically involve large amounts of missing data (the infection process is itself unobserved) and hence we need to find an effective way of estimating the Bayes factors in the presence of this missing information.

Despite the methodological advances in parameter estimation for stochastic models of disease transmission, there is limited off-the-shelf methodology for model selection. In Neal and Roberts (2004) and O'Neill and Marks (2005), model selection for the models of disease transmission have been studied using reversible jump MCMC, whilst Clancy and O'Neill (2007) use rejection sampling to avoid convergence difficulties that are associated with MCMC algorithms. Some more recent examples include Knock and O'Neill (2014), where Bayes factors are computed using path sampling-based algorithms to compare competing models, and O'Neill and Kypraios (2014) who consider the competing models as components of a mixture distribution and then estimate the mixing probabilities which relate to the Bayes factors. Finally, there is growing interest in the use of approximate Bayesian computation methods in model choice problems for epidemic datasets, see for example Toni *et al.* (2009), Lee *et al.* (2015) and Sun *et al.* (2015).

The aim of the current chapter is to demonstrate a straightforward mechanism for estimating the marginal likelihood of models with large amounts of missing data. The idea combines MCMC, importance sampling and filtering in a natural and semi-automatic manner to produce marginal likelihood estimates. The details of the algorithm developed are given Section 4.2. In Section 4.3 we show how the method can be adopted for general applications with household epidemic data and describe the longitudinal *pneumococcal* carriage study motivating this analysis. Simulated data are provided in Section 4.4 to illustrate the implementation, performance and applicability of our algorithm and its comparative performance against a range of alternatives. Finally in Section 4.5 we briefly discuss extensions and limitations of the algorithm.

4.2 Algorithm

4.2.1 The importance sampling estimator

The first observation is that we can rewrite Equation (4.2) as

$$\pi(\mathbf{y}) = \int_{\boldsymbol{\theta}} \pi(\mathbf{y} \mid \boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}, \tag{4.3}$$

where $q(\boldsymbol{\theta})$ denotes a d-dimensional probability density function. We assume that if $\pi(\boldsymbol{\theta}) > 0$ then $q(\boldsymbol{\theta}) > 0$. Then an unbiased estimator, $\widehat{P}_q$ of $\pi(\mathbf{y})$ is obtained by sampling $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(N)}$ from $q(\boldsymbol{\theta})$ and setting

$$\widehat{P}_q = \frac{1}{N} \sum_{i=1}^{N} \pi\left(\mathbf{y} \mid \boldsymbol{\theta}^{(i)}\right) \frac{\pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)}. \tag{4.4}$$

Thus $\widehat{P}_q$ is an importance sampled (see, for example, Ripley, 1987) estimate of $\pi(\mathbf{y})$, and the effectiveness of the estimator given by Equation (4.4) depends upon the variability of $\pi\left(\mathbf{y} \mid \boldsymbol{\theta}^{(i)}\right) \pi\left(\boldsymbol{\theta}^{(i)}\right) / q\left(\boldsymbol{\theta}^{(i)}\right)$.

The optimal choice of $q(\boldsymbol{\theta})$ is $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, the posterior density but if we knew this then $\pi(\mathbf{y})$ would also be known. A simple solution is to use output from an MCMC algorithm to inform the proposal distribution, Clyde *et al.* (2007). For most statistical models the likelihood times the prior is unimodal for sufficiently large number of observations n. In these circumstances, the posterior distribution of $\boldsymbol{\theta}$ is almost always approximately Gaussian with mean $\widehat{\boldsymbol{\theta}}$, the posterior mode, and covariance matrix $\Sigma = -\mathcal{I}(\widehat{\boldsymbol{\theta}})^{-1}$, where $\mathcal{I}(\boldsymbol{\theta})$ denotes the Fisher information evaluated at $\boldsymbol{\theta}$. That is, we have a central limit theorem type behaviour similar to that observed for maximum likelihood estimators as $n \to \infty$. This central limit theorem approximation is implicitly behind the Laplace approximations of integrals used in Tierney and Kadane (1986) and Gelfand and Dey (1994). This underpins the simple suggestion in Clyde *et al.* (2007) of using a multivariate t-distribution as an importance sampling distribution with location and scale parameters estimated from MCMC output. In what follows we found that using a mixture of a multivariate Gaussian distribution, estimated from the MCMC output, and the prior, a “defense mixture” (Hesterberg, 1995) worked well in practice guarding against the proposal density decreasing to 0 faster than the target density.

4.2.2 Missing data

Thus far we have not addressed the issues of computing $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ or dealing with missing data. If $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ is analytically available then the above importance sampling procedure will often work well as long as d is small. However, in many situations $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ is not available. Nonetheless, it is often possible, with the addition of augmented data $\mathbf{x}$, to obtain an analytical expression for $\pi(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta})$. This can then be utilised within an MCMC algorithm to obtain samples from the joint posterior $\pi(\boldsymbol{\theta}, \mathbf{x} \mid \mathbf{y})$. Devising an importance sampling proposal distribution $q(\boldsymbol{\theta}, \mathbf{x})$ approximating $\pi(\boldsymbol{\theta}, \mathbf{x} \mid \mathbf{y})$ will not be practical if $\mathbf{x}$ is high-dimensional, such as in the examples considered in Section 3.6 where $\mathbf{x}$ has elements on the order of thousands. The solution that we propose is to use the marginal MCMC output from $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ to inform the proposal distribution $q(\boldsymbol{\theta})$ in the importance sampling above, and then to separately consider the computation of $\pi(\mathbf{y} \mid \boldsymbol{\theta})$, for which the augmented data $\mathbf{x}$ are required.

For the datasets that are analysed in this chapter there is a temporal structure to the data $\mathbf{y}$ which can be exploited in the estimation of $\pi(\mathbf{y} \mid \boldsymbol{\theta})$. Moreover, filtering methods are particularly well suited to this fixed parameter scenario. In the epidemic model described in Section 4.4, $\mathbf{x}$ represents the unobserved infectious status of individuals with respect to *Streptococcus pneumoniae* carriage and the FFBS algorithm can be used to calculate $\pi(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})$, and hence $\pi(\mathbf{y} \mid \boldsymbol{\theta})$. More details are given in Section 4.3.3.

4.2.3 Variance of the importance sampling estimator

The variance of the importance sampling estimator given in Equation (4.4) is given by:

$$\begin{aligned}\mathrm{Var}(\widehat{P}_q) &= N^{-1} \int_{\boldsymbol{\theta}} \left(\pi(\mathbf{y} \mid \boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} - \pi(\mathbf{y}) \right)^2 q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\ &= N^{-1} \pi(\mathbf{y})^2 \int_{\boldsymbol{\theta}} \left(\frac{\pi(\boldsymbol{\theta} \mid \mathbf{y})}{q(\boldsymbol{\theta})} - 1 \right)^2 q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta},\end{aligned}$$

which highlights that the importance proposal $q(\boldsymbol{\theta})$ should be made to resemble the posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ as closely as possible. As the dimension of $\boldsymbol{\theta}$ increases the variance of the estimator will typically grow due to the curse of dimensionality (see Doucet and Johansen (2011), page 671 for an explanation). For data augmentation problems the dimension of the missing data is often much larger than the number of model parameters, and so if the missing data are treated in the same way as the parameters then this importance sampling approach fails. However if $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ can be calculated, either directly or by using suitable missing data $\mathbf{x}$, then the variance of the importance sampling estimator does not depend on the dimension of the missing data and the importance sampling approach can be applied efficiently, even for large scale data augmentation problems.

If $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ is not available then it must again be estimated, and one can attempt a similar importance sampling approach. Since:

$$\pi(\mathbf{y} \mid \boldsymbol{\theta}) = \int_{\mathbf{x}} \pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) \pi(\mathbf{x} \mid \boldsymbol{\theta}) \, \mathrm{d}\mathbf{x},$$

then the following importance sampling estimator provides an unbiased estimate for $\pi(\mathbf{y} \mid \boldsymbol{\theta})$:

$$\widehat{P}_r = \frac{1}{M} \sum_{j=1}^{M} \pi\left(\mathbf{y} \mid \mathbf{x}^{(j)}, \boldsymbol{\theta}\right) \frac{\pi\left(\mathbf{x}^{(j)} \mid \boldsymbol{\theta}\right)}{r\left(\mathbf{x}^{(j)} \mid \boldsymbol{\theta}\right)},$$

where the $\mathbf{x}^{(j)}$'s are sampled from some proposal distribution $r(\cdot \mid \cdot)$. This is the approach used in pseudo-marginal methods (Beaumont, 2003; Andrieu and Roberts, 2009; McKinley *et al.*, 2014) for estimating the likelihood in the presence of missing data. For estimating the marginal likelihood we can integrate these ideas in one of two ways:

$$\widehat{P}_{rq} = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi\left(\mathbf{y} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(i)}\right) \pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}^{(i)}\right) \pi\left(\boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}^{(i)}\right) q\left(\boldsymbol{\theta}^{(i)}\right)}, \tag{4.5}$$

where $\boldsymbol{\theta}^{(i)} \sim q(\cdot)$ and $\mathbf{x}^{(i)} \sim r\left(\cdot \mid \boldsymbol{\theta}^{(i)}\right)$, or

$$\widetilde{P}_{rq} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\pi\left(\mathbf{y} \mid \mathbf{x}^{(i,j)}, \boldsymbol{\theta}^{(i)}\right) \pi\left(\mathbf{x}^{(i,j)} \mid \boldsymbol{\theta}^{(i)}\right) \pi\left(\boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x}^{(i,j)} \mid \boldsymbol{\theta}^{(i)}\right) q\left(\boldsymbol{\theta}^{(i)}\right)}, \tag{4.6}$$

where $\boldsymbol{\theta}^{(i)} \sim q(\cdot)$ and $\mathbf{x}^{(i,j)} \sim r\left(\cdot \mid \boldsymbol{\theta}^{(i)}\right)$. These are both unbiased and consistent estimators of $\pi(\mathbf{y})$, with $\widehat{P}_{rq}$ the special case of $\widetilde{P}_{rq}$ with $M = 1$. In the Appendix C.1 we show that for a fixed computational effort, the optimal choice (i.e. the one with the lowest variance) is always $\widehat{P}_{rq}$. The variance of Equation (4.5) is given by:

$$\operatorname{Var}\left(\widehat{P}_{rq}\right) = \frac{\pi(\mathbf{y})^2}{N} \int_{\boldsymbol{\theta}} \int_{\mathbf{x}} \left(\frac{\pi\left(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y}\right) \pi\left(\boldsymbol{\theta} \mid \mathbf{y}\right)}{r\left(\mathbf{x} \mid \boldsymbol{\theta}\right) q\left(\boldsymbol{\theta}\right)} - 1 \right)^2 r\left(\mathbf{x} \mid \boldsymbol{\theta}\right) q\left(\boldsymbol{\theta}\right) \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}, \tag{4.7}$$

and this will nevertheless scale with the amount of missing information in the system, which further highlights the need for the proposal distribution for $\mathbf{x}$ to resemble as closely as possible the full conditional $\pi(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})$.

4.3 Epidemic model

4.3.1 *Pneumococcal* carriage study and transmission model

A longitudinal household study of preschool children under 3 years old and all household members was conducted in the United Kingdom from October 2001 to July 2002 (Hussain *et al.*, 2005). The size of the families varied from 2 to 7, although in the most there were 3 or 4 members. All family members were examined for *Streptococcus pneumoniae* carriage using nasopharyngeal swabs once every 4 weeks over a 10-month period. The carriage status of each individual was recorded at each occasion as 1, if a carrier or 0, if a non-carrier.

Following Melegaro *et al.* (2004), we consider a model for transmission of *pneumococcal nasopharyngeal* carriage (Pnc) within a household. At any given time, an individual is assumed to be in either the susceptible non-carrier state 0, or the infectious carrier state 1. The population is divided into two age groups, children under 5 years old and everyone else greater than 5 years (whom for brevity we refer to as ‘adults’), denoted by $i = 1, 2$, respectively. Let $I_1(t)$ and $I_2(t)$ denote the numbers of carrier children and carrier adults in the household at time t. The transition between state 0 and 1 is referred to as an infection and the reverse transition is referred to as clearance. The transition probabilities between states in a short time interval δt are defined for an individual in the age group i:

$$\mathbb{P}(\text{Infection in } (t, t+\delta t]) = 1 - \exp\left\{-\left(\alpha_i + \frac{\beta_{1i}\, I_1(t) + \beta_{2i}\, I_2(t)}{(z-1)^w}\right)\delta t\right\} \tag{4.8}$$

$$\mathbb{P}(\text{Clearance in } (t, t+\delta t]) = 1 - \exp(-\mu_i \cdot \delta t), \tag{4.9}$$

where μ_i and α_i are the clearance and the community acquisition rates respectively for age group i and z is the household size. The rate β_{ij} is the transmission rate from an infected individual in age group i to an uninfected individual in age group j. The term $(z-1)^w$ in Equation (4.8) represents a density correction factor, where w corresponds to the level of density dependence and $(z-1)$ is the number of other family members in a household size z. For example, $w = 1$ represents frequency dependent transmission, where the average number of contacts is equal for each individual in the population. Finally, the probability of infection at the initial swab is assumed to be ν_i for age group i. We refer to this model as $\mathcal{M}_1$.

These definitions allow the carriage within a household to be viewed as a discrete time Markov chain, with time step δt, where the carriage status of each individual depends only on the carriage status of all household members at the previous time point. Because of the dependency between individuals in the same household, a state in the Markov chain consists of the binary vector of states of all of the individuals in the household. The presence of unobserved events, that may have occurred in between swabbing intervals, has been discussed previously by Auranen *et al.* (2000), and must be considered in setting up the model. The approach adopted here to overcome this issue is to use Bayesian data augmentation methods. Model fitting is performed within a Bayesian framework using an MCMC algorithm, imputing the unobserved carriage states of each household.

Let $O^p \subseteq \{1, 2, \dots, T\}$ denote the set of prescheduled observation times of household $p = 1, 2, \dots, P$, and let $U^p = \{1, 2, \dots, T\} \setminus O^p$ denote the unobserved times. Let $\mathbf{y}_t^p$ be the binary vector of carriage states for individuals in household p at observation time t. The observed longitudinal data $\mathbf{y} = \left[\mathbf{y}_t^p\right]_{t \in O^p;\, p=1,\dots,P}$ consists of the household carriage statuses $\mathbf{y}_t^p$ at the observation times. Similarly let $\mathbf{x}_t^p$ be the corresponding latent carriage status of household p at time $t \in U^p$, and form the corresponding missing data matrix $\mathbf{x} = \left[\mathbf{x}_t^p\right]_{t \in U^p;\, p=1,\dots,P}$. Let $\boldsymbol{\theta}$ denote the vector of model parameters, including the rates of acquiring and clearing carriage, the density correction w and the initial probabilities of carriage.

4.3.2 Markov chain Monte Carlo algorithm

In the Bayesian approach, the missing data is represented as a nuisance parameter and inferred from the observed data like any other parameter. The joint posterior density of the latent carriage states $\mathbf{x}$, and the model parameters $\boldsymbol{\theta}$ can be factorized as:

$$\pi(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \prod_{p=1}^{P} \prod_{t=1}^{T} \pi\left(\mathbf{z}_t^p \mid \mathbf{z}_{t-1}^p, \boldsymbol{\theta}\right),$$

where $\mathbf{z}_t^p$ equals $\mathbf{y}_t^p$ if $t \in O^p$; $\mathbf{x}_t^p$ if $t \in U^p$ and $\emptyset$ if $t = 0$. This factorisation is based on the assumption that conditionally on the model parameters, the carriage process is assumed to be independent across households.

In order to simulate from the posterior distribution, we construct an MCMC algorithm that employs both Gibbs and Metropolis-Hastings updates. The main emphasis is on sampling the unobserved carriage process $\mathbf{x}$, which we do using a

Gibbs step via the forward filtering backward sampling algorithm, as described in Section 3.4.1. In the first part of this algorithm, recursive filtering equations are used to calculate $\mathbb{P}\left(\mathbf{x}_t^p \mid \mathbf{z}_{t+1}^p, \mathbf{y}_{O^p \cap \{1:t\}}^p, \boldsymbol{\theta}\right)$ for each $t \in U^p$ working forwards in time. The second part then works backwards through time, simulating $\mathbf{x}_t^p$ from these conditionals, starting with $t = \max(U^p)$ and ending with $t = \min(U^p)$. The model parameters ν_1 and ν_2 are updated using Gibbs updates and the remaining parameters are updated jointly using an adaptive Metropolis-Hastings random walk proposal (Roberts and Rosenthal, 2009).

4.3.3 Marginal likelihood estimation via importance sampling

The availability of the full conditional distribution of the missing data $\pi(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})$ from the FFBS algorithm allows the missing data component $\mathbf{x}$ to be updated using a Gibb's step in the MCMC algorithm. This full conditional can be exploited further in the estimation of the marginal likelihood. We require $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ in order to form the importance sampling estimator in Equation (4.4). Using Bayes' Theorem we can rewrite this as

$$\pi(\mathbf{y} \mid \boldsymbol{\theta}) = \frac{\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})\, \pi(\mathbf{x} \mid \boldsymbol{\theta})}{\pi(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})} = \frac{\pi(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta})}{\pi(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})}, \tag{4.10}$$

for any $\mathbf{x}$ such that $\pi(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}) > 0$. Therefore evaluation of $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ at the point $\boldsymbol{\theta}$ can be done by evaluating the right-hand-side of Equation (4.10) with any suitable $\mathbf{x}$. A suitable $\mathbf{x}$ is guaranteed if it is sampled from the full conditional distribution $\mathbf{x} \mid (\mathbf{y}, \boldsymbol{\theta})$.

Our approach proceeds as follows. In step 1 we use MCMC to obtain samples from the joint posterior of $\boldsymbol{\theta}$ and $\mathbf{x}$. In step 2 we fit a multivariate normal distribution to the posterior samples for $\boldsymbol{\theta}$ only, and use it to construct a normalised proposal density $q(\boldsymbol{\theta})$. In step 3, we obtain N samples from $q(\boldsymbol{\theta})$ and for each sample $\boldsymbol{\theta}^{(i)}$ we obtain a corresponding sample for the missing data $\mathbf{x}^{(i)}$ using the forward filtering backward sampling algorithm. We then use these samples to calculate the importance sampling estimator of the marginal likelihood:

$$\widehat{P}_q(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi\left(\mathbf{y}, \mathbf{x}^{(i)} \mid \boldsymbol{\theta}^{(i)}\right)}{\pi\left(\mathbf{x}^{(i)} \mid \mathbf{y}, \boldsymbol{\theta}^{(i)}\right)} \frac{\pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)}. \tag{4.11}$$

The choice of $q(\boldsymbol{\theta})$ is important for the accuracy and computational efficiency of the importance sampling approach. As discussed in Section 4.2.3, we want $q(\boldsymbol{\theta})$ to be a good approximation of $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ but with heavier tails to ensure that the variance of

$\widehat{P}_q$ is small. We therefore investigate a range of proposals distributions based on a fitted multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ based on the MCMC output. These include drawing $\boldsymbol{\theta}$ from $\text{IS}_{N_j} : \mathcal{N}(\boldsymbol{\mu}, j\boldsymbol{\Sigma})$ $(j = 1, 2, 3)$, a multivariate Normal distribution with different variances; $\text{IS}_{t_v} : t_v(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $(v = 4, 6, 8)$, a multivariate Student's t distribution with v degrees of freedom, mean $\boldsymbol{\mu}$ and covariance matrix $\frac{v}{v-2}\boldsymbol{\Sigma}$ (if $v > 2$) and $\text{IS}_{\text{mix}} : q(\boldsymbol{\theta}) = 0.95 \times \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0.05 \times \pi(\boldsymbol{\theta})$ (mixture of a multivariate Normal density and the prior).

4.4 Simulation studies

4.4.1 Marginal likelihood estimation

We consider the problem of estimating the marginal likelihood under the model introduced in Section 4.3, using the methods described above. These estimators were evaluated on synthetic data analogous to the real data in Melegaro *et al.* (2004). More specifically, the parameter values were based on the maximum likelihood estimates from the analysis of Pnc data; parameters were chosen to be $\alpha_1 = 0.012$, $\alpha_2 = 0.004$, $\beta_{11} = 0.047$, $\beta_{12} = 0.005$, $\beta_{21} = 0.106$, $\beta_{22} = 0.048$, $\mu_1 = 0.020$, $\mu_2 = 0.053$, $w = 1.184$, $\nu_1 = 0.425$ and $\nu_2 = 0.095$. We set the time-interval $\delta t = 7$. Only complete family transitions, where the infection state of all household members was known on two consecutive observations, were used previously by Melegaro *et al.* (2004) (51% of the full dataset). Although our approach could easily handle the missing data, for comparability we match the number of complete transitions by family size and number of adults to generate our data set; a total of 66 families comprising 260 individuals including 94 children under 5 years. The simulations were designed so that real and simulated datasets have the same sampling times. The hidden variable $\mathbf{x}$ consists of 1650 $\mathbf{x}_t^p$'s, comprising 6500 unobserved binary variables in total.

We compare the proposed importance sampling approach for estimating the marginal likelihood (based on the 7 proposal densities) with bridge sampling (Meng and Wong, 1996) (using the importance samples from IS_{mix}), harmonic mean (Newton and Raftery, 1994), Chib's method (Chib, 1995; Chib and Jeliazkov, 2001) and the power posteriors method (Friel and Pettitt, 2008). Details of the computation of these estimators are given in Section C.2 of the Appendix. To compare the different methods on a fair basis, we chose to dedicate equivalent amounts of computational effort for estimation of the log marginal likelihood, instead of fixing the total number of samples.

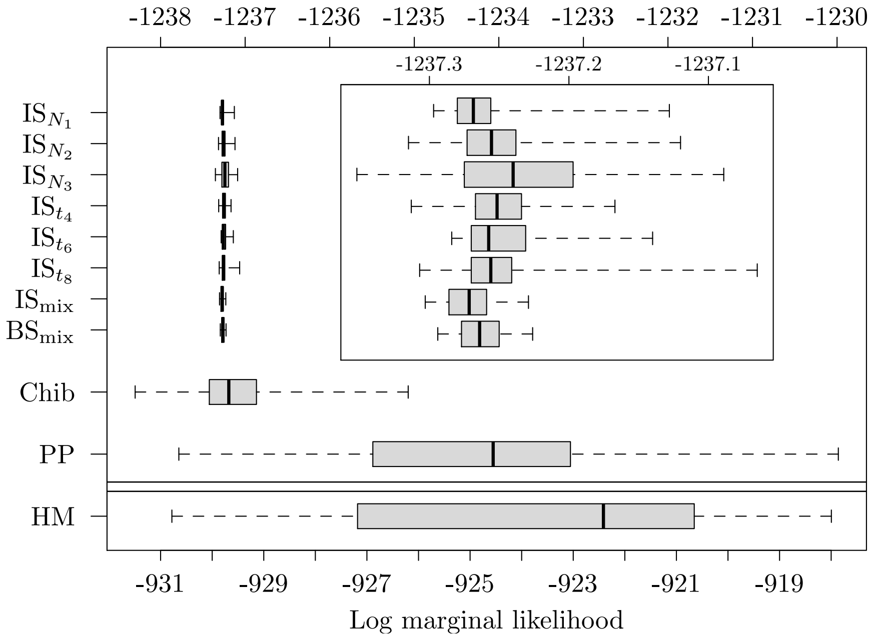
Implementation details are given as follows. The construction of the importance density was based on 25,000 MCMC samples after a burn-in of 5,000, obtained from the MCMC sampler described in Section 4.3.2. These posterior samples were used to estimate the reference parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for a multivariate Student's t or normal proposal density. The marginal likelihood estimate was then based on 25,000 importance sampling draws from the obtained proposal density $q(\boldsymbol{\theta})$, using the estimator in Equation (4.11). To produce the bridge sampling estimate, the 25,000 samples from IS_{mix} were combined with 250 thinned samples from the MCMC. In order to apply Chib's methods, the same posterior samples were used for computing the high posterior density point. The log marginal was estimated by generating 22,000 draws in each complete and reduced MCMC run, with the first 2,000 draws removed as burn-in. Harmonic mean analysis was based on 50,000 posterior samples, following a 3,000 iteration burn-in. For the power posterior method, it was necessary to specify the temperature scheme and a pilot analysis (not counted in the computation cost) was used to choose 20 partitions on the unit interval. The MCMC sampler was run for 2,650 iterations for each temperature in the descending series, omitting the first 650 as burn-in, finishing with 2,650 samples at $t = 0$ (the prior).

Each procedure was repeated 50 times to provide an empirical Monte Carlo estimate of the variation in each approach. We also vary the total running time in order to investigate the effect of this on the accuracy of the marginal likelihood estimates. For each analysis method we used the same priors: Ga(0.01,0.01) for the density factor w; Beta(1,1) for the initial probabilities of infection ν_1 and ν_2 and Ga(1,1) for the remaining parameters.

Figure 4.1 shows the variability of the eleven marginal likelihood estimators. Except for the harmonic mean, all the methods appear to have produced consistent estimates of the marginal likelihood. Chib's method produced better estimates of the marginal likelihood than the power posterior method, which is more computationally expensive than the other methods and therefore uses a small number of MCMC samples at each temperature, leading to large uncertainty. However as seen in Figure 4.1, the bridge sampling and the importance sampling methods offer significant improvements in precision over the other methods. Moreover, increasing the number of MCMC samples, led to a decrease in the Monte Carlo standard errors of order $\mathcal{O}(\sqrt{n})$, see Table 4.1, indicating that the variances of the corresponding estimators are finite.

The success of the importance sampling approach is not surprising since it explores the posterior distribution of parameters more efficiently than the other methods due to the independence of the samples drawn from the proposal density.

Figure 4.1: Variability of the log marginal likelihood estimates for model $\mathcal{M}_1$ over 50 replicates for each the methods. Each of the methods have roughly the same computational cost.



More surprisingly we were unable to use the bridge sampling technique to improve substantially on the standard errors, which dropped from 0.0196 for $\mathrm{IS}_{\mathrm{mix}}$ to 0.0179 for $\mathrm{BS}_{\mathrm{mix}}$. The bridge sampling estimator attempts to combine information from the MCMC and importance samples, however the optimal estimator is derived assuming that independent samples from the posterior were available, which we approached by applying a thinning of 100 to the samples. With low levels of thinning (results shown in Figure C.2 of the Appendix) we found that bridge sampling actually increased the standard error of the marginal likelihood estimate.

On the basis of this example, the lowest variance importance sampling estimator was obtained using the proposal density $\mathrm{IS}_{\mathrm{mix}}$ – a mixture of the prior and the normal fitted to the posterior samples. Therefore, from now on we use this proposal density when estimating the log marginal likelihood via importance sampling.

4.4.2 Model comparison

In this section, we apply the marginal likelihood estimation approaches to the problem of Bayesian model choice. We focus on their ability to distinguish between biologically motivated hypotheses concerning the dynamics of Pnc transmission. In particular we compare their performance against the established technique of re-

Table 4.1: Monte Carlo (MC) standard errors of log marginal likelihood estimates for different number of Markov chain samples. Standard errors are given across 50 replicates for each of the methods. Each of the methods have roughly the same computational cost.

Method	**MC samples**	**MC error**	**MC samples**	**MC error**	**MC samples**	**MC error**
IS_{N_1}	10000	0.053	25000	0.033	50000	0.025
IS_{N_2}	10000	0.064	25000	0.036	50000	0.025
IS_{N_3}	10000	0.107	25000	0.061	50000	0.042
IS_{t_4}	10000	0.037	25000	0.025	50000	0.020
IS_{t_6}	10000	0.064	25000	0.034	50000	0.023
IS_{t_8}	10000	0.034	25000	0.035	50000	0.024
IS_{mix}	10000	0.030	25000	0.020	50000	0.012
BS_{mix}	10000	0.029	25000	0.018	50000	0.011
Chib	8000	0.736	20000	0.486	40000	0.312
PP	20×1600	2.906	20×2150	1.936	20×3200	1.547
HM	37000	5.548	50000	5.331	72000	4.850

versible jump Markov chain Monte Carlo and then demonstrate that the importance sampling approach can solve problems that are extremely challenging with RJMCMC. We show that using our approach it is possible to answer epidemiologically important questions such as whether there is heterogeneity in transmission rates and if household size is related to transmission. Finally we consider how the accuracy of the importance sampling approach is affected by the amount of missing data.

4.4.2.1 Heterogeneity in community acquisition rates

Suppose that we wish to evaluate the evidence in favour of the community acquisition rates being equal for adults and children, in the hope of developing a more parsimonious model. We call the model described in Section 4.3, in which children have community acquisition rate α_1 and adults have rate α_2, model $\mathcal{M}_1$. The nested model, in which $\alpha_1 = \alpha_2$ is called $\mathcal{M}_2$. We generated realistic simulated datasets from each of these models and then used importance sampling, bridge sampling, Chib's method, power posteriors, the harmonic mean and reversible jump MCMC to estimate the Bayes factor in favour of $\mathcal{M}_1$, denoted by B_{12}. As before, we used approximately the same computational effort for each of these approaches. For $\mathcal{M}_1$ we assumed $\alpha_1 = 0.012$ and $\alpha_2 = 0.004$, whilst for $\mathcal{M}_2$ we assumed $\alpha_1 = \alpha_2 = 0.008$.

Details of the RJMCMC algorithm for selecting between models $\mathcal{M}_1$ and

$\mathcal{M}_2$ are given in Section C.2.5 of the Appendix. As before, the MCMC samples used for estimating Bayes factors with RJMCMC were designed to be comparable in computational effort with the other methods. Therefore, the RJMCMC chain was allowed a 30,000 burn-in followed by 76,000 samples. When the evidence is strongly in favour of one model, the RJMCMC will not move between models very often and can provide poor estimates of the Bayes factor. A variant of the method, called RJMCMC corrected (RJcor), can tackle this issue by assigning higher prior probability to the model that is visited less often. This probability is estimated as $\pi(k) = 1 - \widehat{\pi}(k \mid \mathbf{y})$, where $\widehat{\pi}(k \mid \mathbf{y})$ is obtained from a pilot run of RJMCMC with initial $\pi(k) = 0.5$, for $k = 1, 2$. For RJcor we did 30,000 pilot iterations and then another 76,000 iterations, of which 30,000 were discarded as a burn in.

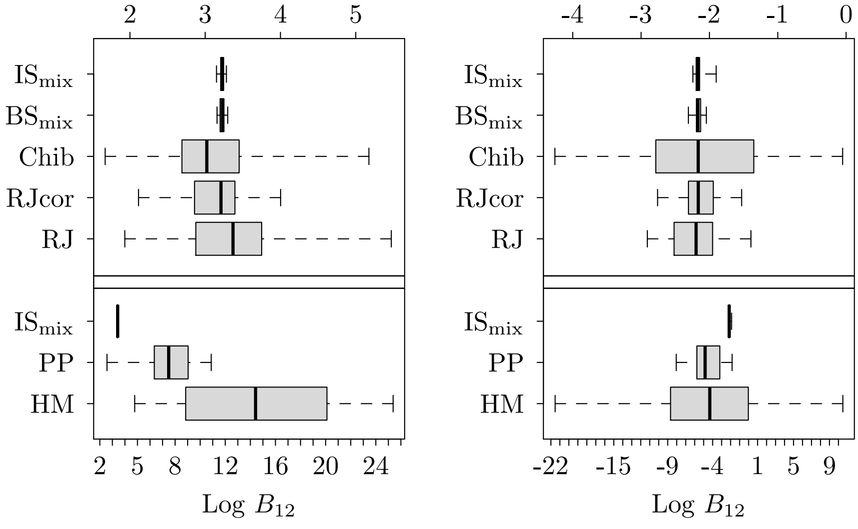
Figure 4.2 provides a graphical representation of the variability in $\log(B_{12})$ over 50 repeats of each Monte Carlo approach. The plot highlights that the estimators based on importance sampling and bridge sampling were the most accurate in both scenarios. The left panel of Figure 4.2 gives results for data generated from $\mathcal{M}_1$. Importance sampling, bridge sampling, Chib and RJ methods lead to similar estimates, whereas power posterior and harmonic mean overestimated the log Bayes factor. Moreover, RJcor produced slightly more accurate estimates of the log Bayes factor than *vanilla* RJMCMC. All methods selected the correct model, with largest variation from the harmonic mean estimator. In the right panel of Figure 4.2, the results use data generated from model $\mathcal{M}_2$. Due to the huge variance in $\log(B_{12})$, the harmonic mean sometimes favoured the wrong model. Although the remaining methods correctly identified the true model, the importance and bridge sampling methods again produced the most precise estimates of the Bayes factor; the standard errors provided by the two methods are almost identical.

Figure 4.3 demonstrates the log Bayes factor in favour of $\mathcal{M}_1$ as a function of computation time using data generated from $\mathcal{M}_1$. The importance sampling estimator (in blue) converges much more rapidly than the other estimators, showing very tight credible intervals. Chib's method (in green) and corrected RJMCMC (in red) appear to converge to the same value, but more slowly and have wider CIs. The power posterior method gradually approaches the consensus estimate, requiring significantly more samples to stabilise. The harmonic mean estimator was heavily unstable and also provided much wider credible intervals than the other methods.

4.4.2.2 Heterogeneity in household transmission rates

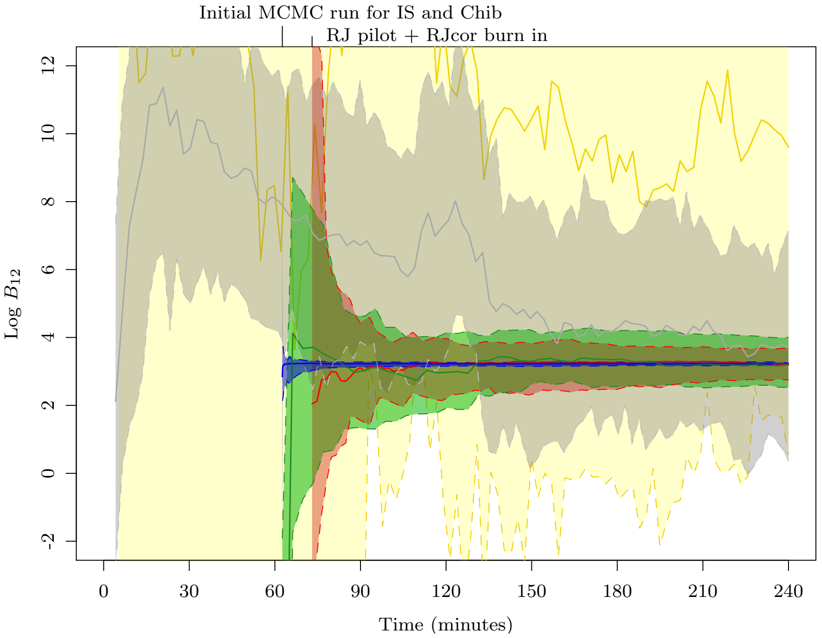
We wish to evaluate whether or not there is heterogeneity in the household transmission rates. More precisely, we wish to compare the full model $\mathcal{M}_1$ with the

Figure 4.2: Variability of the log Bayes factor estimates based on 50 Monte Carlo repeats for the importance sampling method with mixture proposals ($\mathrm{IS}_{\mathrm{mix}}$), bridge sampling method with mixture proposals ($\mathrm{BS}_{\mathrm{mix}}$), Chib's method, reversible jump MCMC (RJ), corrected reversible jump MCMC (RJcor), power posteriors (PP) and harmonic mean (HM) methods.



(a) Data simulated from model $\mathcal{M}_1$. (b) Data simulated from model $\mathcal{M}_2$.

Figure 4.3: Evolution of log Bayes factor estimates in favour of model $\mathcal{M}_1$. The solid lines corresponds to the median and the shaded areas give the 95% credible intervals, estimated from 50 replicates. Yellow represents the harmonic mean method, grey is for the power posterior, red and green correspond to RJMCMC corrected and Chib's methods respectively and blue represents the importance sampling approach with the mixture proposals.



special case in which the within-household acquisition rates are identical between the two age groups, i.e. $\beta_{11} = \beta_{12} = \beta_{21} = \beta_{22} = \beta$ (say), which we call model $\mathcal{M}_3$. This kind of question is extremely challenging to answer using reversible jump methodology because it is difficult to move efficiently between models when this involves a large change in dimension. Again we generated two datasets, one from $\mathcal{M}_1$ using the parameters given in Section 4.4.1, and one from $\mathcal{M}_3$ with $\beta = 0.0515$, the average of $\beta_{11}, \beta_{12}, \beta_{21}$ and β_{22}. For both datasets, we calculated Bayes factors using importance sampling, bridge sampling, Chib's method, power posteriors and the harmonic mean. Our objective was to check that the correct model was chosen by the Bayes factors criterion in this setting.

Table 4.2 presents the marginal likelihood estimates and the corresponding Monte Carlo standard errors for each method, where bold entries show the preferred model. The importance sampling, bridge sampling, Chib and power posterior methods all agreed and were able to discriminate the true model. The estimates of the log marginal likelihoods are similar within Monte Carlo error, with importance and bridge sampling being the most precise. As was previously observed in Section 4.4.1, the harmonic mean overestimated the log marginal likelihoods and yielded inaccurate results, favouring the wrong model in both scenarios.

Table 4.2: Bayes factors and log marginal likelihoods of the main and reduced models for the two simulation designs. The Monte Carlo standard errors over 50 replicates are shown in parentheses. Entries in bold show the selected model for each method and simulation design.

Simulation design	**Method**	**Log marginal of model $\mathcal{M}_1$**	**Log marginal of model $\mathcal{M}_3$**	**logB_{13}**
Data from $\mathcal{M}_1$	IS$_{\text{mix}}$	**-1267.102** (0.018)	-1268.843 (0.020)	1.742 (0.031)
	BS$_{\text{mix}}$	**-1267.104** (0.018)	-1268.844 (0.019)	1.744 (0.029)
	Chib	**-1266.999** (0.261)	-1268.075 (0.619)	1.190 (0.729)
	PP	**-1262.957** (1.926)	-1266.150 (2.107)	3.215 (2.465)
	HM	-931.320 (3.882)	**-929.168** (5.444)	-3.562 (6.507)
Data from $\mathcal{M}_3$	IS$_{\text{mix}}$	-1512.107 (0.011)	**-1505.058** (0.015)	-7.048 (0.019)
	BS$_{\text{mix}}$	-1512.101 (0.011)	**-1505.060** (0.013)	-7.042 (0.021)
	Chib	-1512.110 (0.326)	**-1505.021** (0.290)	-7.156 (0.445)
	PP	-1509.138 (2.003)	**-1500.616** (2.089)	-8.833 (2.495)
	HM	**-1184.755** (5.150)	-1195.668 (6.252)	9.273 (7.552)

4.4.2.3 Density-dependence in within-household transmission

Melegaro *et al.* (2004) investigated the relationship between transmission rates and household size via the density correction factor $(z-1)^w$ in the transmission rates, Equation (4.8), where z is the household size. Since their confidence interval for w included 1 they were unable to determine whether transmission increased ($w < 1$) or decreased ($w > 1$) with household size. Moreover, the value $w = 1$ corresponds to frequency dependent transmission, where the average number of contacts is the same irrespective of household size. We wish to determine whether frequency dependent transmission ($w = 1$, which we call model $\mathcal{M}_4$) could be identified from the data.

Bayesian model comparison problems of this kind often suffer from Lindley's paradox (Robert, 2001), where the choice of prior for w in the more complex model has undue influence on the resulting Bayes factor. To reduce (but not remove) the impact of Lindley's paradox we consider two priors for w in $\mathcal{M}_1$: Ga(1,1) (referred as the local prior) and the inverse moment prior for $\log w$ (referred as the non local prior), with densities respectively given by:

$$\pi_L(w) = \frac{b^a w^{a-1} e^{-wb}}{\Gamma(a)}, \quad a = 1, b = 1,$$

$$\pi_{NL}(w) = \frac{\rho\tau^{\upsilon/2}}{w\,\Gamma(\upsilon/2\rho)}\big(\log(w)\big)^{-(\upsilon+1)} \exp\left[-\left\{\frac{\big(\log(w)\big)^2}{\tau}\right\}^{-\rho}\right],$$

with $\rho = 1$, $\upsilon = 1$ and $\tau = 0.173$ (for more details see Johnson and Rossell, 2010). The density functions of the two priors are shown in Figure 4.4. The figure illustrates the fact that the non local prior has density zero at $w = 1$.

To determine if evidence in favour or against $\mathcal{M}_4$ could be determined from the study of Melegaro *et al.* (2004) we simulated datasets of equivalent size with values for w from 0.5 through to 2, increasing by 0.1 each time. For each value of w we obtained an estimate of the posterior probability of $\mathcal{M}_1$ along with its standard error, based on 100 simulated datasets. Results are shown in Figure 4.5. For values of w close to 1, the non local prior provided on average stronger evidence in favour of the simple model even though model $\mathcal{M}_1$ was technically the correct model. For values of w within the interval [0.6, 1.4] both priors supported $\mathcal{M}_4$, but only the non local prior provided positive support for $\mathcal{M}_4$. Whereas when w went from 1.5 to 2, both priors favoured $\mathcal{M}_1$, with the non local prior providing equal or higher posterior probability in favour of the correct model than the alternative local prior. Melegaro *et al.* (2004) estimated $w = 1.18$ and in this region we expect weak support for frequency dependent transmission, model $\mathcal{M}_4$.

Figure 4.4: Prior distributions on the density correction factor w.

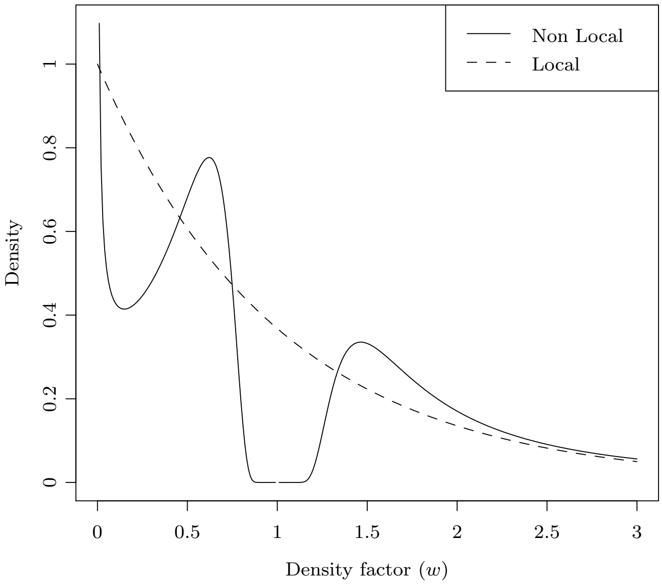
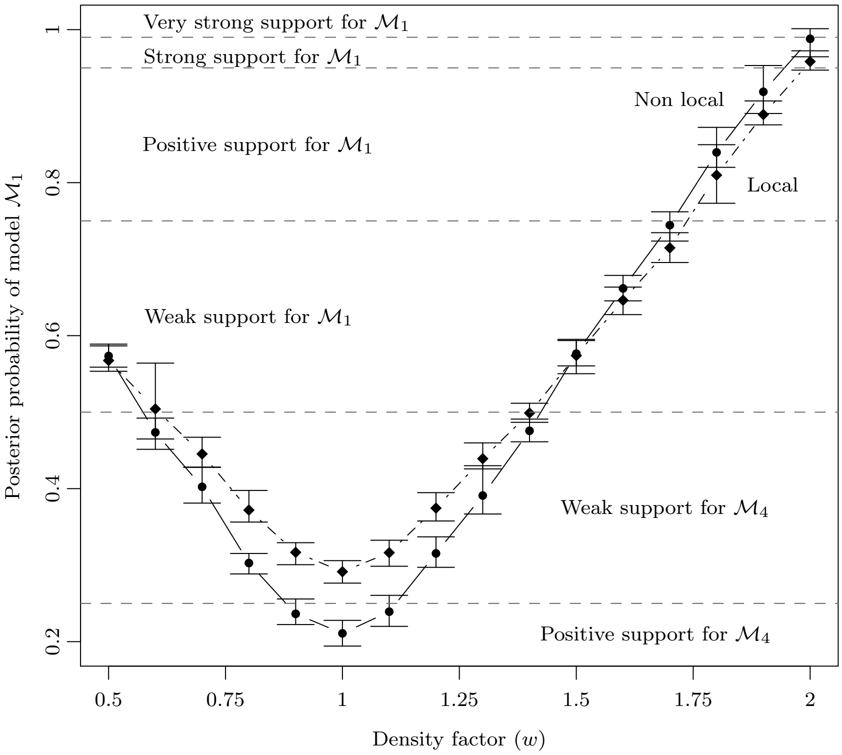


Figure 4.5: Posterior probability of the full model using two different prior specifications; the local prior (– · – · –) and the non local prior (———). Error bars represent the Monte Carlo standard error based on 100 simulations.



4.4.2.4 Amount of missing data

In this section we are interested in assessing the accuracy and efficiency of the proposed method as a function of the total number of hidden states. One way to vary the amount of missing data without diluting the information content of the dataset is to vary the time interval δt. The larger δt is, the smaller the number of hidden states that need to be imputed. For example when $\delta t = 1$, 60840 hidden states need to be imputed, whereas we have only 3900 when $\delta t = 10$. For $\delta t = 1, 2, \ldots, 10$, we generated 10 synthetic datasets according to $\mathcal{M}_1$. For each dataset we fitted 10 different models, one for each possible value of δt, and calculated the log marginal likelihood.

For brevity, Figure 4.6 presents results only for the data generated by $\delta t = 1, 5, 10$ and methods IS and Chib. In all three cases, the log marginal likelihood curves are peaked at the true value of δt, the one used to create the data (Figure 4.6). The marginal likelihood estimates from Chib's method are also maximised at true values, but since the standard errors are much higher, more samples would be required to distinguish between the competing models.

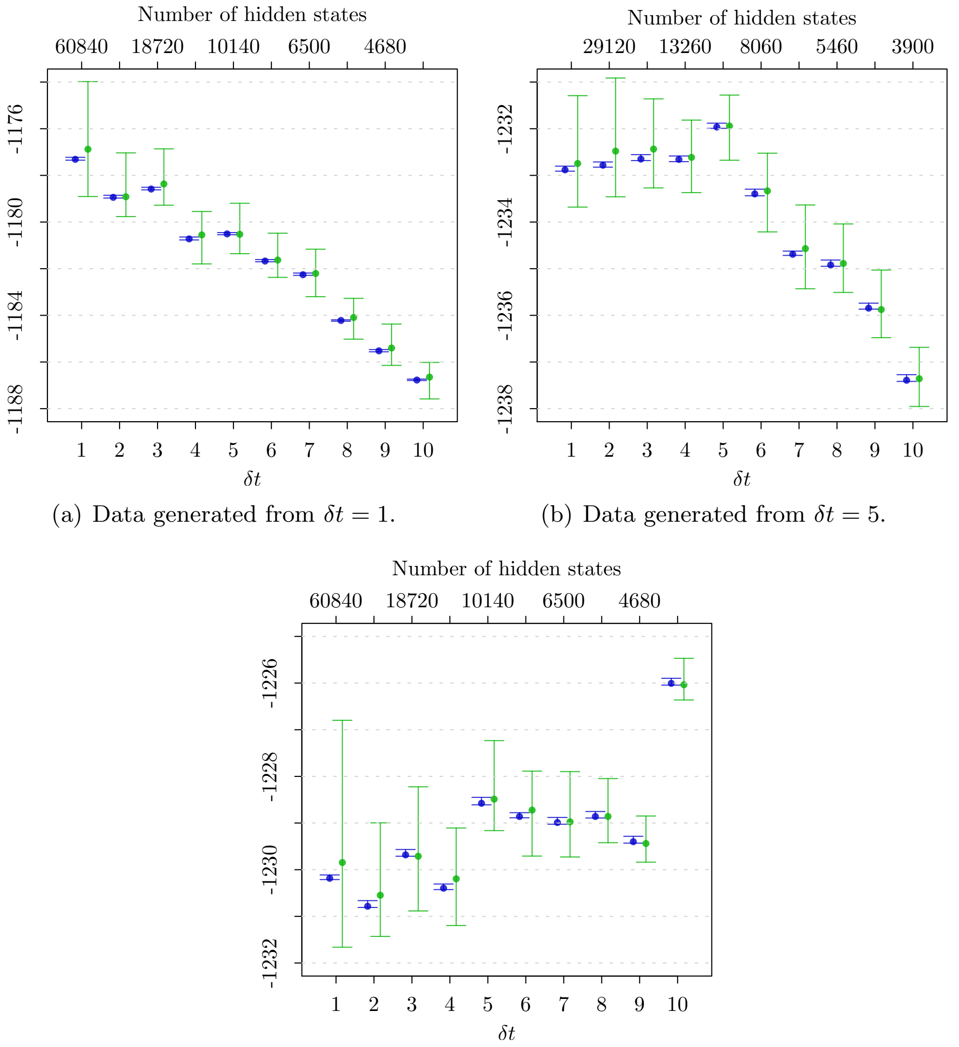
Figure 4.7 shows how the Monte Carlo standard errors in Chib's method and the importance sampling method increase as a function of the total number of hidden states. The graph shows that the Monte Carlo standard errors from the importance sampling method appear very stable as the dimensionality of the hidden states is increased.

4.5 Conclusions

In this chapter, we have introduced a simple three-stage algorithm for efficiently estimating the marginal likelihood in applications with several missing data. The key components are an MCMC algorithm for obtaining samples from the posterior distribution, $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, an approximating distribution $q(\boldsymbol{\theta})$ to sample from and an effective estimate of the likelihood $\pi(\mathbf{y} \mid \boldsymbol{\theta})$. The first observation is whilst an MCMC algorithm will often be relatively straightforward to construct, alternative methods for sampling from the posterior distribution could be equally considered. Moreover, it is not important if a sample from an approximate posterior distribution (for example, Monte Carlo within Metropolis, O'Neill *et al.*, 2000) is used since all that is required for computation of the marginal likelihood is to be able to make a reasonable choice of $q(\cdot)$.

Furthermore the importance sampling and the associated estimation of the likelihood is trivially parallelisable which can be utilised to speed up implementation.

Figure 4.6: Sensitivity to δt, the time interval. The figure shows the variability of the log marginal likelihood estimation when using data generated by (a) $\delta t = 1$, (b) $\delta t = 5$ and (c) $\delta t = 10$. Blue and green colors represent the importance sampling and Chib's methods respectively.



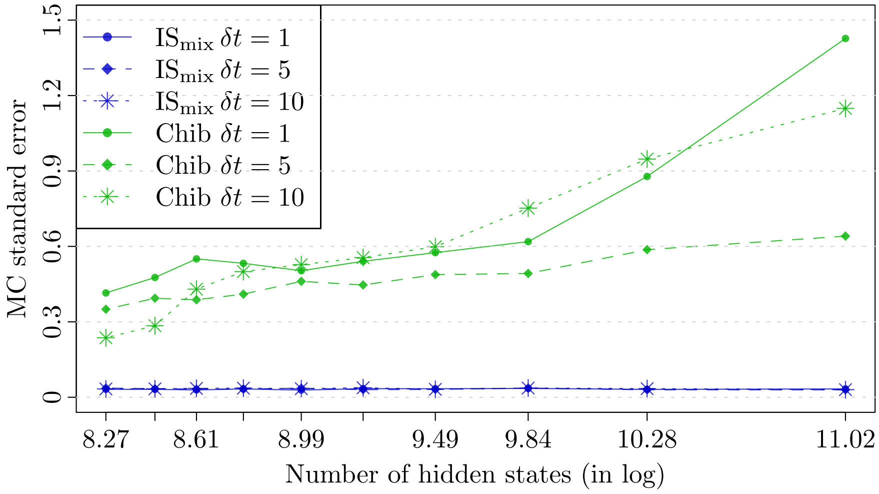
(a) Data generated from $\delta t = 1$.

(b) Data generated from $\delta t = 5$.

(c) Data generated from $\delta t = 10$.

In cases where the likelihood can easily be computed the algorithm becomes a simple add-on to MCMC to compute the marginal likelihood. An additional advantage of the IS approach is that even if we consider a further model, we can apply the method relatively quickly as we only need to calculate the log marginal likelihood for this extra model, without recomputing the marginal likelihoods of the other models. In contrast, in a RJMCMC setting such analysis might be computationally expensive since there is a need to run the algorithm including the new model and design new

Figure 4.7: Monte Carlo standard error of the proposed importance sampling and Chib's methods for different number of hidden states and values of time interval δt.



efficient jump proposals.

The key limitation to using this approach is effective estimation of the likelihood $\pi(\mathbf{y} \mid \boldsymbol{\theta})$ in cases where it is not analytically tractable. For the epidemic examples considered in this chapter we have been able to exploit the temporal nature of the data to use filtering methods to estimate $\pi(\mathbf{y} \mid \boldsymbol{\theta})$. However, the method is applicable to longitudinal data more generally. Another application where the method that we describe can be of significant utility is time series data in which case particle filtering can be used to obtain estimates of $\pi(\mathbf{y} \mid \boldsymbol{\theta})$, see Touloupou *et al.* (2015) for more details.

We have applied the methodology to data simulated from a household epidemic model, comparing our algorithm to existing methods for computing the marginal likelihood. For a fixed computational budget, the IS method provides estimates of similar quality to the bridge sampling approach and outperforms the remaining of the methods that we consider. Further, it was demonstrated how these estimates can be used to deliver epidemiological questions of interest using model comparison to choose between several competing hypotheses. Finally, we found that estimates of the marginal likelihood using IS were very stable (in terms of variability) as a function of the total amount of missing data as opposed to Chib's method with which it was compared.

Chapter 5

Bayesian Model Selection For Evaluation Of Epidemiological Hypotheses: The Epidemiology of Escherichia Coli O157:H7 In Cattle

5.1 Introduction

In epidemics, it is often challenging to design and validate an appropriate model for the data under study, particularly in diseases for which relatively little is known regarding the transmission process. Therefore, models are usually chosen based on expert knowledge, authors' judgements or information regarding a similar disease. However, these decisions may lead to mathematical models that are either very simple or very complex to represent the data properly. For example, early epidemic models assumed that individuals of a community mix homogeneously that is, have the same infectivity and susceptibility to a disease (Bailey, 1975). Such a simplification may not be plausible for all studies, e.g. when additional heterogeneities could arise due to characteristics of individuals such as age or social contacts. One possible way to address this complication is to compare several transmission models in order to further our understanding regarding the dynamics of a disease.

In this chapter, we demonstrate the usefulness of the statistical tools for model comparison that were developed in Chapter 4, by uncovering new insights into the transmission dynamics of *E. coli* O157:H7 in cattle. In our application, each one of the competing models represents an important hypothesis regarding the epidemiology of the disease, for which relatively little is known. These questions of interest are either biologically motivated or naturally arise due to the experimental design of datasets 1 and 2, to which we apply our methods. When possible, we validate our findings by applying the reversible jump MCMC algorithm, additionally to our importance sampling approach.

The first hypothesis that we are interested in testing is whether cattle develop immunity to *E. coli* O157:H7 over time. This question relates to the distribution of the colonisation period. We therefore compare two distributions, where in the first the probability of recovery is constant over time whereas in the second, this probability depends on how long an individual has been carrying the disease. We then assess the extent to which pen-specific factors are important in the disease spread. For example, we consider factors such as the size and the geographic location, which may differ depending on the pen and could affect its members risk of colonisation. As a third study, we investigate potential routes of infection between pens that are located within close distance of each other. In the datasets that we consider, such routes include shared waterers, feed bunks and boundaries and we aim to identify which one (if any) mostly contributes to the spread of the disease.

The rest of this chapter is structured as follows. In Sections 5.2, 5.3 and 5.4 we apply model selection procedures to investigate the three hypotheses concerning the transmission of *E. coli* O157:H7 that we introduced above. In each of these sections, we explicitly state the competing models, provide the algorithm implementation details and interpret the results of each analysis. Finally, in Section 5.5 we summarise our findings and propose some directions for future research.

5.2 Investigating the distribution of the colonisation period

5.2.1 Competing models

The colonisation period of disease is defined as the number of consecutive days that an individual remains colonised by it. The majority of research on stochastic epidemic models assumes that the colonisation period is a Markov random variable that is, the probability of an individual being cleared is constant over time. This choice is not always epidemiologically motivated but is nevertheless employed because it makes the statistical analysis easier. A common choice is the Geometric distribution or its continuous time analogue the Exponential distribution, which both have the memoryless property. For many diseases however, it is more plausible to assume that the probability of recovery varies according to how long an animal has been infected. In such cases a Negative Binomial distribution or the Gamma distribution for continuous time models are more well suited for the colonisation period.

In this section we compare the Geometric (described in Section 3.2) with the Negative Binomial model in order to determine which of the two is more appropriate for *E. coli* O157:H7 based on our dataset 1. Therefore, two candidates models are presented, corresponding to different assumptions concerning the distribution of the colonisation period Z,

Model $\mathcal{M}_1$: the colonisation period is assumed to be Geometrically distributed with mean m, $Z \sim \text{Geom}(m)$.

Model $\mathcal{M}_2$: Z follows a Negative Binomial distribution with mean m and dispersion parameter κ, $Z \sim \text{NB}(m, \kappa)$.

For clarity of notation, in both models we have used parametrization in terms of the mean and therefore the two corresponding probability mass functions for the Geometric and Negative Binomial model, are given respectively by:

$$\pi_1(\zeta; m) := \mathbb{P}_1(Z = \zeta) = \left(\frac{m-1}{m}\right)^{\zeta-1} \times \frac{1}{m},$$

$$\pi_2(\zeta; \kappa, m) := \mathbb{P}_2(Z = \zeta) = \left(\frac{\kappa}{\kappa+m-1}\right)^{\kappa} \frac{\Gamma(\kappa+\zeta-1)}{(\zeta-1)!\,\Gamma(\kappa)} \left(\frac{m-1}{\kappa+m-1}\right)^{\zeta-1},$$

which are supported on the integers $\zeta \in \{1, 2, \ldots\}$, where $m \geq 1$ is the mean duration of the disease and $\kappa > 0$ is the dispersion parameter. Note that Z denotes a discrete random variable that represents the duration time of a completed colonisation, and therefore, under the Negative Binomial model, if an individual is found colonised at the first day of the study and we assume that they have just acquired the disease then the duration will appear shorter than it is and will bias our estimate. Similar bias is introduced if we assume that an individual that is colonised on the last day of the study cleared the disease on that day. In order to correct for such left and right censoring, we propose the use of the size-biased sampling density function (Rao, 1965), given by:

$$\pi_2^s(\zeta; \kappa, m) = \frac{\zeta \times \pi_2(\zeta; \kappa, m)}{\mathbb{E}(Z)}.$$

Under this assumption, the probability that an individual is colonised at the beginning of the study and remains so for ζ^* days is given by,

$$\mathbb{P}_2(Z^* = \zeta^*) \quad = \quad \sum_{\zeta=\zeta^*}^{\infty} \mathbb{P}_2(Z^* = \zeta^* \mid Z = \zeta) \times \pi_2^s(\zeta; \kappa, m)$$

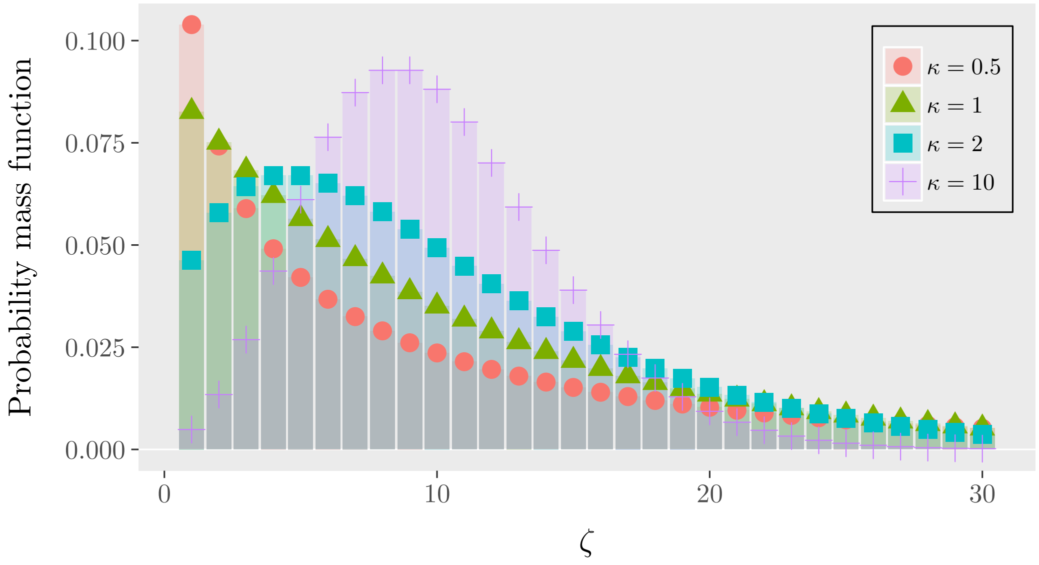
$$
\begin{aligned}
&= \sum_{\zeta=\zeta^*}^{\infty} \frac{1}{\zeta} \times \frac{\zeta \times \pi_2(\zeta;\kappa,m)}{\mathbb{E}(Z)} = \sum_{\zeta=\zeta^*}^{\infty} \frac{\mathbb{P}_2(Z=\zeta)}{m} \\
&= \frac{\mathbb{P}_2(Z \geq \zeta^*)}{m},
\end{aligned}
$$

where following Cox *et al.* (2005) we assume that the conditional distribution of the observed colonisation time Z^* is uniform over $\{0, 1, \ldots, \zeta\}$ given the complete duration of colonisation $Z = \zeta$. When an individual is found to be in the colonised state at time $t = T$ (the last day of the study), the starting time of the colonisation is known but not the time that the individual is cleared. Therefore, in this case we have:

$$\mathbb{P}_2(Z^* = \zeta^*) = \mathbb{P}_2(Z \geq \zeta^*).$$

We remark that the Negative Binomial model includes the Geometric model as a special case for $\kappa = 1$. The mean of both distributions is m, but the variances differ between the two models. In particular, the variance is $m(m-1)$ for the Geometric distribution whereas it is $(m-1) + \frac{(m-1)^2}{\kappa}$ for the Negative Binomial. Hence, the Negative Binomial allows for a more flexible shape as shown in Figure 5.1 for $\kappa = 0.5, 1, 2, 10$ and $m = 10$.

Figure 5.1: Probability mass functions of the Negative Binomial distribution with mean 10 and four different values of the dispersion parameter $\kappa = 0.5, 1, 2, 10$. Lower values of κ correspond to more over-dispersed distributions.



5.2.2 Implementation details

In Section 3.4 we have shown how posterior samples can be obtained for the SIS model under a Geometrically distributed colonisation period. Our MCMC algorithm takes advantage of the fact that the full conditional distribution of the unobserved colonisation states for each pen p (which we denote $\pi_H\left(\mathbf{X}^p \mid \mathbf{Y}^p, \boldsymbol{\theta}_1\right)$, $\boldsymbol{\theta}_1 = (\alpha_1, \beta_1, m_1, \nu_1, \theta_{R_1}, \theta_{F_1})$ being the parameters of model $\mathcal{M}_1$) can be found in a closed form and thus we can use a Gibbs step to update $\mathbf{X}_1^p$, for $p = 1, 2, \ldots, P$. However, under the Negative Binomial model, the history of each individual must be represented explicitly in the model and therefore it is hard to compute the full conditional of $\mathbf{X}_2^p$, the hidden process of colonisation under $\mathcal{M}_2$.

Instead of having a Gibbs step, we use π_H as a proposal distribution in a Metropolis-Hastings step. In particular, we consider an independence sampler where for each $p = 1, 2, \ldots, P$ we propose $\mathbf{X}^{p*} \sim \pi_H$ and the move is accepted with probability:

$$\min\left(1, \frac{\pi_H\left(\mathbf{X}_2^p \mid \mathbf{Y}^p, \kappa = 1, \boldsymbol{\theta}_2^{-\kappa}\right)}{\pi_H\left(\mathbf{X}^{p*} \mid \mathbf{Y}^p, \kappa = 1, \boldsymbol{\theta}_2^{-\kappa}\right)} \times \frac{\pi_2\left(\mathbf{X}^{p*}, \boldsymbol{\theta}_2 \mid \mathbf{Y}^p\right)}{\pi_2\left(\mathbf{X}_2^p, \boldsymbol{\theta}_2 \mid \mathbf{Y}^p\right)}\right),$$

where $\boldsymbol{\theta}_2 = (\alpha_2, \beta_2, m_2, \nu_2, \theta_{R_2}, \theta_{F_2}, \kappa)$, $\boldsymbol{\theta}_2^{-\kappa} = \boldsymbol{\theta}_2 \setminus \{\kappa\}$ and $\pi_2\left(\mathbf{X}_2^p, \boldsymbol{\theta}_2 \mid \mathbf{Y}^p\right)$ is the posterior distribution of model $\mathcal{M}_2$ given in Section D.1 of the Appendix. The approximate full conditional essentially assumes that $\kappa = 1$ and therefore the performance of the algorithm depends on how close the true value of κ is to 1. Parameters ν, θ_R and θ_F are updated as in the Geometric model with Gibbs steps, and the remaining parameters α, β, m and κ are updated jointly with Hamiltonian Monte Carlo (see Appendix D.1 for derivative expressions).

Model comparison is carried out with the importance sampling approach that was introduced in Chapter 4. For comparison, we also apply the RJMCMC method which has been broadly used in the context of model selection for epidemic models. We first give details of the RJMCMC procedure using the notation of Section 1.2.4.7. The probability of jumping from model k to k', where $k, k' \in \{1, 2\}$ is fixed at 0.5. A priori, we assume that $\pi(k = 1) = \pi(k = 2) = 0.5$. There are four types of move: (1) Geometric to Geometric, (2) Geometric to Negative Binomial, (3) Negative Binomial to Geometric, and (4) Negative Binomial to Negative Binomial. Only moves 2 and 3 are trans-dimensional; for moves 1 and 4 we use MCMC updates described above. An important aspect regarding the efficiency of RJMCMC is the specification of the proposal distribution and the transformation function $g_{1,2}$. Here we consider three approaches for constructing the dimension increasing proposal.

The first proposal is suggested by Hastie and Green (2012). Suppose the Markov chain is in the Geometric model; then, jumping to the Negative Binomial distribution (move 2) can be done by generating u from a Normal density, $u \sim \mathcal{N}(0, \sigma^2)$ with σ fixed but well chosen in order to achieve good mixing between models. For the transformation function they use $\boldsymbol{\psi}_2 = (\mathbf{X}_2, \alpha_2, \beta_2, m_2, \mu_2, \theta_{R_2}, \theta_{F_2}, \kappa) = g_{1,2}(\boldsymbol{\psi}_1, u) = \big(\boldsymbol{\psi}_1, \lambda \exp(u)\big)$, with λ fixed and $\boldsymbol{\psi}_1 = (\mathbf{X}_1, \alpha_1, \beta_1, m_1, \nu_1, \theta_{R_1}, \theta_{F_1})$. With this choice we keep $\mathbf{X}, \alpha, \beta, m, \mu, \theta_R, \theta_F$ fixed when switching between models and the parameter κ is a Lognormal random variable centered at λ. The Jacobian of the transformation is $\lambda \exp(u)$.

For the reverse move 3 we go back to the Geometric space and therefore there is no need to generate a random variable since the parameter space is reduced. It sufficient to use the inverse transformation of $g_{1,2}$, that is $g_{2,1} = g_{1,2}^{-1}$, to transform the variables back to the Geometric model. This is achieved by taking $(\boldsymbol{\psi}_1, u) = g_{2,1}(\boldsymbol{\psi}_2) = (\mathbf{X}_2, \alpha_2, \beta_2, m_2, \mu_2, \theta_{R_2}, \theta_{F_2}, \log \frac{\kappa}{\lambda})$. The Jacobian of the transformation is then $\frac{1}{\kappa}$.

Therefore, for the move from model $\mathcal{M}_1$ to model $\mathcal{M}_2$, the probability of accepting the jump is given by $\min(1, A_{12})$ where:

$$A_{12} = \frac{\pi(k=2, \boldsymbol{\psi}_2 \mid \mathbf{Y})}{\pi(k=1, \boldsymbol{\psi}_1 \mid \mathbf{Y})} \left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{u^2}{2\sigma^2} \right] \right\}^{-1} \lambda \exp(u),$$

whereas for the reciprocal move from $\mathcal{M}_2$ to $\mathcal{M}_1$, the probability of accepting the jump is given by $\min(1, A_{21})$ where:

$$A_{21} = \frac{\pi(k=1, \boldsymbol{\psi}_1 \mid \mathbf{Y})}{\pi(k=2, \boldsymbol{\psi}_2 \mid \mathbf{Y})} \left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\big(\log(\kappa/\lambda)\big)^2}{2\sigma^2} \right] \right\} \frac{1}{\kappa}.$$

A second option is to keep parameters that appear in both models fixed, and set the additional parameter $\kappa = u$, where $u \sim \mathrm{Exp}(\xi)$. This is equivalent to taking the matching functions $g_{1,2}(\boldsymbol{\psi}_1, u) = (\boldsymbol{\psi}_1, u)$ and $g_{2,1}(\boldsymbol{\psi}_2) = \boldsymbol{\psi}_2$, as the identity functions. This is appropriate since the Geometric model is nested within the Negative Binomial model and its parameters have identical interpretation with the corresponding parameters of the larger model. In this manner, the acceptance probability for the move from model $\mathcal{M}_1$ to model $\mathcal{M}_2$ is calculated according to Equation (1.4) in Section 1.2.4.7, with the acceptance probability of the reverse move given by the reciprocal of this value. The Jacobian term in the acceptance probability equals one.

Finally, following Dellaportas and Forster (1999), we start by performing a pilot MCMC run for the Negative Binomial model in order to obtain estimates of the marginal posterior moments of the additional parameter κ. For RJMCMC, a Gamma proposal distribution is then employed to generate $u \sim \text{Ga}(a, b)$; this distribution is considered as the third proposal. The parameters a and b can be calculated by equating the mean and the variance of the Gamma distribution with the estimated mean ($\hat{\kappa}$) and variance ($\hat{\sigma}_\kappa^2$) of κ from the output of the pilot MCMC run. Hence:

$$\hat{\kappa} = \frac{a}{b}, \quad \hat{\sigma}_\kappa^2 = \frac{a}{b^2},$$

which leads to

$$a = \frac{\hat{\kappa}^2}{\hat{\sigma}_\kappa^2}, \quad b = \frac{\hat{\kappa}}{\hat{\sigma}_\kappa^2}. \tag{5.1}$$

For the IS technique we use Equation (4.5) and therefore need to specify the proposal densities q and r. Let $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ be the mean and the variance-covariance matrix of the parameters (either $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$), as estimated from the MCMC output consisting of 20,000 draws (after 6,000 burn-in) from the posterior. Following our findings in Chapter 4, we choose $q(\boldsymbol{\theta}_k) = 0.95 \times \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + 0.05 \times \pi(\boldsymbol{\theta}_k)$ for both candidate models. For the Geometric model, the construction of the proposal density $r(\mathbf{X}_1 \mid \cdot)$ is done according to the FFBS algorithm, as described in Section 4.3.3. For the Negative Binomial model, we cannot sample directly from the full conditional distribution of the unobserved colonisation states as we previously explained. Instead, we use the same proposal as for the Geometric model, based on the assumption that $\kappa = 1$, and then the importance weights correct for the fact that this is an approximation.

For all implementations in this section, we assume mutually independent prior distributions for the model parameters, and specifically that $\alpha, \beta \sim \text{Ga}(1, 1)$, $m - 1 \sim \text{Ga}(0.01, 0.01)$, $\nu, \theta_R, \theta_F \sim \text{Beta}(1, 1)$. For the additional parameter κ in the Negative Binomial model we do a sensitivity analysis to explore its effect on the Bayes factor. We run RJMCMC with proposals 1-2 for 240,000 iterations with the first 50,000 discarded as burn-in. For proposal 3 the pilot run consists of 26,000 MCMC draws and since we keep the computation budget fixed, we allow RJMCMC to run for 190,000 iterations of which the first 50,000 are accounted as burn-in. The pilot run is also used to construct the importance sampling density q and we then draw 100,000 samples from the proposal densities to obtain the IS estimates.

5.2.3 Simulation studies

In this section we evaluate the ability of the proposed importance sampling and RJMCMC methods to distinguish between the Geometric and Negative Binomial models with simulated data. We generate datasets under the Negative Binomial model with true parameter values set to $\alpha = 0.009$, $\beta = 0.01$, $m = 9$, $\nu = 0.1$, $\theta_R = 0.8$ and $\theta_F = 0.5$, whereas for the dispersion parameter we use 8 different scenarios where κ is successively set to $0.25, 0.5, \dots, 2.0$. Note that in scenario 4 where $\kappa = 1$ we essentially create data under the Geometric model. The total number of pens, individuals per pen and samples obtained are chosen equal to the corresponding attributes of dataset 1, for all 8 scenarios. As discussed in Section 4.4.2.3, in order to diminish Lindley's paradox effect we use the inverse moment (non local) prior for $\log \kappa$ which leads to the following prior density for $\kappa > 0$:

$$\pi(\kappa) = \frac{\rho \tau^{\upsilon/2}}{\kappa \, \Gamma(\upsilon/2\rho)} \big(\log(\kappa)\big)^{-(\upsilon+1)} \exp\left[-\left\{ \frac{\big(\log(\kappa)\big)^2}{\tau} \right\}^{-\rho} \right], \tag{5.2}$$

with $\rho = 1$, $\upsilon = 1$ and $\tau = 0.16$. Note that the non local prior has density 0 at $\kappa = 1$.

In order to compare the results of the IS analysis with those of RJMCMC, we present the posterior model probabilities of the Negative Binomial model in Table 5.1. To eliminate biases due to the simulated data, we create and analyse 40 datasets for each value of κ and report the median posterior probability over these replicates. Comparing the two methods, we see that they give nearly identical estimates of the posterior probability of model $\mathcal{M}_2$ for all scenarios considered. IS outperforms the 3 RJMCMC samplers in terms of variability of the estimates, except from the cases where $\kappa = 0.25$ and $\kappa = 2.00$. This result could be attributed to the proposal distribution of $\mathbf{X}$ being less efficient as κ moves further away from 1. Additionally, when $\kappa = 2$, the posterior probability assigned to model $\mathcal{M}_2$ by IS is slightly lower compared to RJMCMC. However, in the two extreme cases we are uncertain whether RJMCMC exhibits good mixing or if it is stuck in one model. In cases where the algorithm remains at a model for a long period, one might use the corrected RJMCMC to ameliorate this problem. Of the 3 RJMCMC proposals, we find that proposal 3 is the one that provides the less accurate estimates, due to the fact that MCMC runs for less iterations. Further, the Normal proposal results in the lowest variability as well as the highest probability of moving between the 2 models as can be seen in Tables 5.1 and 5.2, respectively. Nevertheless, our results support that posterior model probabilities are robust to different choices of proposal distributions in RJMCMC.

In scenarios 1 and 8, all algorithms yield a very high posterior probability that the true underlying distribution is the Negative Binomial (Table 5.1). This can directly be linked to the true value of κ (0.25 and 2.00 for these scenarios) which contrasts with the assumption of the Geometric model where $\kappa = 1$. In fact, in a run of RJMCMC using the third proposal under scenario 1, it is virtually impossible to accept the Geometric model against the Negative Binomial model. As expected, the evidence in favour of the Negative Binomial model decreases in the remaining scenarios and in particular for 3-6 we observe that the Negative Binomial appears less favourable than the Geometric model. The lowest probabilities among the 8 scenarios are obtained when $\kappa = 1$, where we correctly obtain strong evidence in favour of the simple model.

In terms of parameter estimation, all the simulation studies show that the MCMC algorithm is able to provide estimates of good quality. In Table 5.3 we report posterior summaries of parameters κ and m which we calculate over 40 replicates in each one of the 8 scenarios. Overall, we conclude that it is possible to recover the parameters since in all setups the 95% intervals contain the true values that were used to generate the data.

Finally, we carry out a sensitivity analysis to assess the effect that the prior of the additional parameter κ has on posterior model probabilities. Additional to the non local prior, we also consider the $\text{Ga}(1, 1)$ and $\text{Ga}(0.1, 0.1)$ priors, the latter being the least informative of the three. We use the same κ as in scenarios 1-8 and for every value we repeat estimation 40 times with each one of the three priors. The median (over the replicates) posterior probabilities of the Negative Binomial model $\mathcal{M}_2$ are shown in Figure 5.2. As expected, the results differ depending on the prior. The uninformative $\text{Ga}(0.1, 0.1)$ prior tends to favour the simple Geometric model

Table 5.1: Estimated median posterior probability (standard deviation) of the Negative Binomial model, $\mathcal{M}_2$, based on 40 replicates under eight different epidemic scenarios.

True κ	**Method**			
	IS	**RJMCM Normal**	**RJMCM Exp**	**RJMCM Gamma**
0.25	0.987 (0.021)	0.997 (0.003)	0.993 (0.001)	1.000 (0.000)
0.50	0.867 (0.021)	0.869 (0.059)	0.870 (0.065)	0.871 (0.033)
0.75	0.404 (0.009)	0.417 (0.028)	0.410 (0.028)	0.395 (0.044)
1.00	0.117 (0.002)	0.101 (0.011)	0.118 (0.014)	0.116 (0.058)
1.25	0.255 (0.005)	0.253 (0.010)	0.257 (0.013)	0.215 (0.091)
1.50	0.444 (0.010)	0.443 (0.015)	0.446 (0.021)	0.458 (0.023)
1.75	0.782 (0.019)	0.778 (0.020)	0.781 (0.020)	0.783 (0.022)
2.00	0.955 (0.021)	0.975 (0.005)	0.974 (0.010)	0.983 (0.011)

Table 5.2: Median probability of a move between the Geometric and Negative Binomial models based on 40 replicates under eight different epidemic scenarios. Estimates are multiplied by 100.

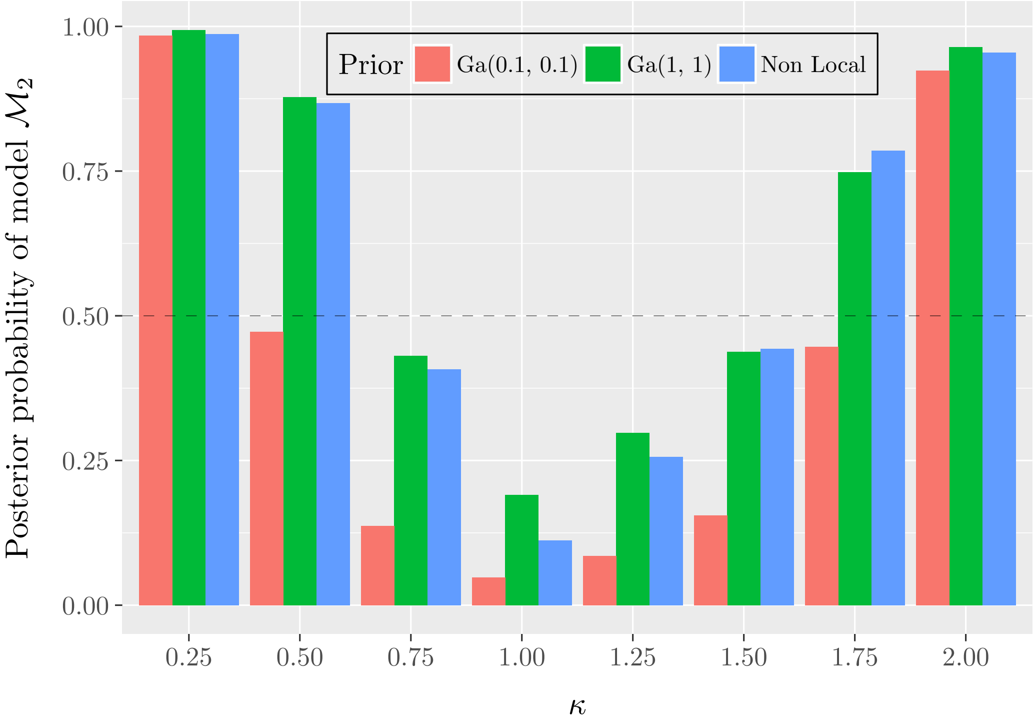
True κ	Method		
	RJMCM Normal	RJMCM Exp	RJMCM Gamma
0.25	0.044	0.144	0.000
0.50	1.304	1.708	2.064
0.75	7.493	5.004	7.047
1.00	7.119	4.769	6.028
1.25	11.556	6.790	9.526
1.50	12.464	5.936	9.113
1.75	4.500	2.625	4.013
2.00	2.224	0.556	0.911

Table 5.3: Posterior summaries of parameters κ and m based on 40 replicates under eight different epidemic scenarios.

Scenario	Dispersion κ				Mean duration m			
	2.5%	$\hat{\kappa}$	97.5%	True	2.5%	$\hat{m}$	97.5%	True
1	0.147	0.263	0.388	0.25	8.855	9.086	9.337	9.00
2	0.377	0.507	0.647	0.50	8.584	8.845	9.127	9.00
3	0.633	0.755	0.906	0.75	8.757	9.018	9.324	9.00
4	0.929	1.046	1.165	1.00	8.888	9.124	9.357	9.00
5	1.092	1.211	1.329	1.25	8.645	8.891	9.156	9.00
6	1.340	1.490	1.612	1.50	8.840	9.142	9.386	9.00
7	1.642	1.789	1.924	1.75	8.735	9.030	9.279	9.00
8	1.937	2.112	2.229	2.00	8.464	8.815	9.048	9.00

in all scenarios except when $\kappa = 0.25$ and $\kappa = 2.00$. The other 2 priors generally provide similar results with the exception of scenarios 3, 4 and 5 when κ is close to 1. In these cases, the non local prior gives more evidence for model $\mathcal{M}_1$ as it assigns less mass to values of κ near 1 and exactly 0 at 1.

Figure 5.2: Sensitivity to the prior distribution on the additional parameter κ. Results are based on 40 replicates under eight different epidemic scenarios. We use 3 different priors: the Ga(0.01, 0.01), the Ga(1,1) and the non local prior given by Equation (5.2).

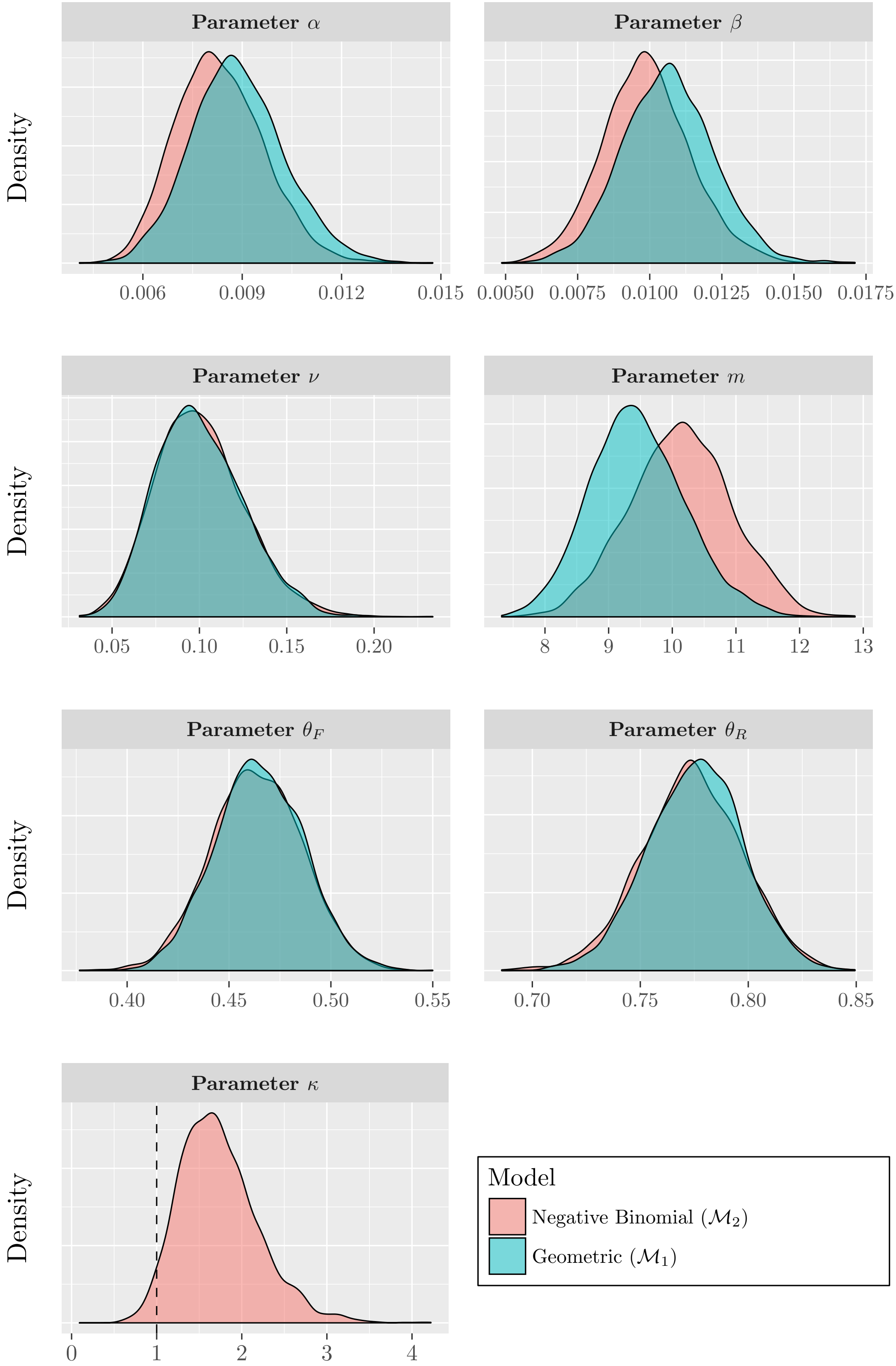


5.2.4 Results

We first present parameter estimates for the *E. coli* O157:H7 dataset 1 under the Geometric and Negative Binomial models and then we consider the problem of determining which of the two best fits the data. When comparing the marginal posterior densities of the common parameters obtained by the two analyses (see Figure 5.3), we observe that these are in close agreement, with the main difference being in the mean colonisation period. In particular, the parameter m is lower in the Geometric model, leading to higher estimates of transmission parameters α and β. The tail is longer on the Geometric distribution compared with Negative Binomial ($\kappa > 1$) and this may be causing the mean colonisation period to be underestimated. The estimated posterior median of κ under the Negative Binomial model is 1.681 and the corresponding 95% credible interval is equal to [0.975, 2.758]. This might weakly suggest that the Geometric model is not plausible for this data.

To formally test between the two models, the IS and RJMCMC methods are

Figure 5.3: Marginal posterior densities of the parameter when analysing *E. coli* O157:H7 dataset 1 under the Geometric and Negative Binomial models. The green and red densities correspond to model $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively.



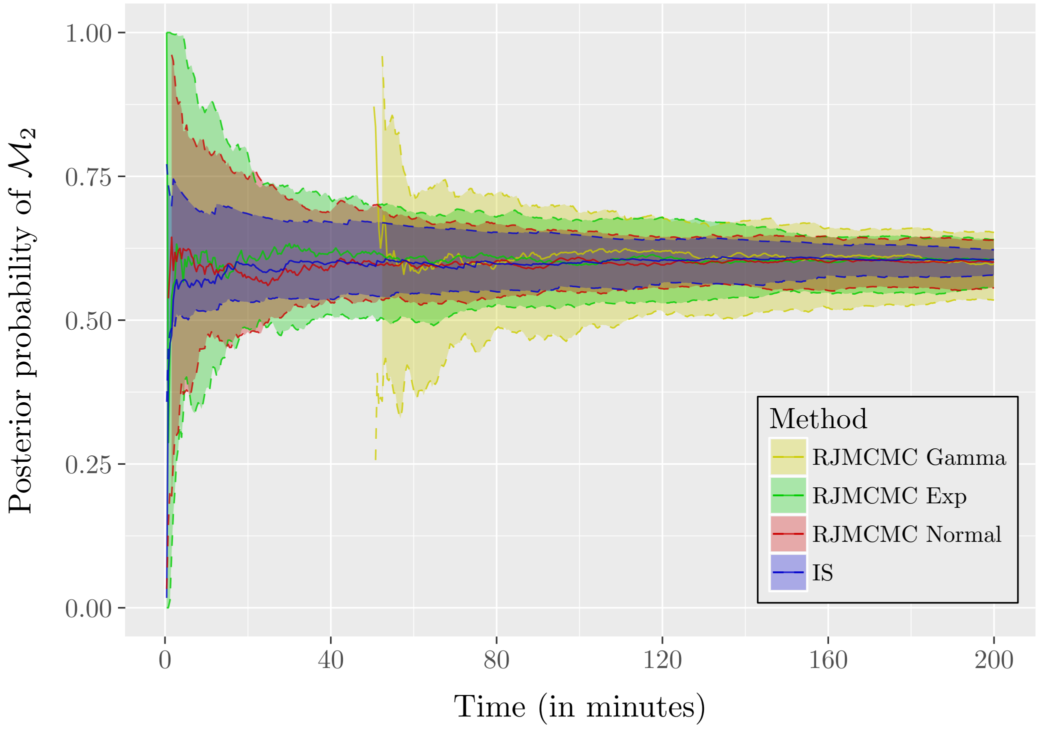
implemented. For the RJMCMC, we use the three different dimension increasing proposal densities presented in Section 5.2.2. The parameter σ of the Normal proposal (proposal 1) is set to 0.5 and the parameter λ in the matching function for an across-model jump is set to 1.5. Furthermore, the parameter ξ of the Exponential proposal (proposal 2) distribution is fixed to 1. Lastly, the resulting values of a and b for the Gamma density (proposal 3) are obtained from the pilot run of the Negative Binomial model, based on Equation (5.1). All the Markov chains are initiated with the Geometric model.

Median posterior probabilities of the two models, along with the logarithm of the estimated Bayes factor in favour of the Negative Binomial model are presented in Table 5.4, using the non local prior for κ. Note that we have applied all methods 40 times to the dataset to gain an estimate of the variability of the quantities presented. The results of IS and RJMCMC are almost identical and support the Negative Binomial model with a median posterior probability of 0.60. However this evidence is not overwhelming. Similar to our simulation studies, we find that IS estimates are slightly less variable compared to those obtained by RJMCMC. We also plot the evolution of the posterior probability of model $\mathcal{M}_2$ (Figure 5.4) and see that after a reasonably long run, all methods converge towards the same value.

Table 5.4: Estimated Log Bayes factors, posterior model probabilities and the probability of a move between the two models for the real dataset 1. $\mathcal{M}_1$: Geometric model, $\mathcal{M}_2$: Negative Binomial model, $\mathbb{P}(\mathcal{M}_1 \mid \mathbf{y})$: posterior probability of model $\mathcal{M}_1$ and $\mathbb{P}(\mathcal{M}_2 \mid \mathbf{y})$: posterior probability of model $\mathcal{M}_2$, $\mathbb{P}$(move): probability of a move between the two models.

Methods	$\mathbb{P}(\mathcal{M}_1 \mid \mathbf{y})$	$\mathbb{P}(\mathcal{M}_2 \mid \mathbf{y})$	**Log Bayes factor**	$\mathbb{P}$**(move)**
IS	0.394 [0.378, 0.422]	0.606 [0.578, 0.622]	0.429 [0.345, 0.532]	—
RJMCMC Normal	0.398 [0.362, 0.444]	0.602 [0.556, 0.638]	0.413 [0.183, 0.523]	0.155 [0.151, 0.158]
RJMCMC Exp	0.394 [0.360, 0.443]	0.606 [0.557, 0.640]	0.433 [0.229, 0.577]	0.105 [0.103, 0.111]
RJMCMC Gamma	0.402 [0.347, 0.465]	0.598 [0.535, 0.653]	0.396 [0.274, 0.691]	0.142 [0.093, 0.151]

Figure 5.4: Evolution of the posterior probability of the Negative Binomial model using four different methods applied on *E. coli* O157:H7 dataset 1.



5.2.5 Discussion

In summary, the simulation analyses in this section illustrate that both IS and RJMCMC methods perform very well in distinguishing between the Geometric and Negative Binomial model for the colonisation period. Moreover, we have seen that IS generally produces more precise estimates compared to RJMCMC, especially in cases where the dispersion parameter κ of the Negative Binomial model is close to 1. Further, the posterior model probabilities obtained by RJMCMC were independent of the trans-dimensional proposal that was used for the algorithm. Finally, a sensitivity analysis has suggested that the prior on the dispersion parameter κ can possibly affect the Bayes factor estimation.

Valuable conclusions have been drawn after comparing the Geometric and Negative Binomial model on dataset 1. The results from the analysis show that the non-Markovian Negative Binomial model is better supported by the data than the simpler, Markovian Geometric model. Specifically, there is weak evidence, as suggested by a Bayes factor of 1.54 obtained by IS. This result proves some evidence in support of the hypothesis that the probability of an individual leaving the colonised state is not constant over time, and in particular that the probability of recovery grows as the colonisation period increases. The biological interpretation of the finding is that cattle may develop an immune response to rid themselves of *E. coli* bacteria. Thus, in subsequent analyses of dataset 1 we assume a Negative Binomial distributed colonisation period.

5.3 Investigating heterogeneity in colonisation rates among pens

5.3.1 Competing models

In this section we consider the model with Negative Binomial distribution for the colonisation periods as suggested by our analysis in Section 5.2. Further, we extend this model to incorporate heterogeneity in the spread of the disease through the study population. More precisely, we assume that the rates of avoiding colonisation from within-pen and from external sources vary between pens.

Looking at the map with the locations of all the pens in the first study (Figure 2.1), we can see that the pens are divided into two sets according to their size and their location. The North pen set consists of 12 pens with dimensions 6m × 17m (area 102m^2) and the South pen set includes the remaining 8 pens that have dimensions 6m × 37m (area 222m^2). It is likely that pens with different area and location are subject to different levels of risk of *E. coli* O157:H7 infection. Therefore, we allow the external and within-pen transmission rates α and β, respectively, to differ for South and North pens. The corresponding parameters are indexed by s for South pens and n for North pens.

To investigate the difference between the two sets of pens, we define the model with two risk levels to this data set, with parameters (α_s, β_s) corresponding to individuals housed in South pens and (α_n, β_n) corresponding to individuals housed in North pens. Hence, we assume that the overall avoidance probability for individual c housed in pen p on a given day t, is given by

$$\begin{aligned}&\mathbb{P}\left(X_t^{[c,\,p]} = 0 \mid X_{t-1}^{[c,\,p]} = 0\right)\\ &\quad = \exp\left(- \alpha_s \mathbb{1}_{\{p\in\mathcal{S}\}} - \alpha_n \mathbb{1}_{\{p\in\mathcal{N}\}} - \beta_s\, \mathbb{1}_{\{p\in\mathcal{S}\}} \sum_{c'=1}^{n_c^p} x_{t-1}^{[c',p]} - \beta_n\, \mathbb{1}_{\{p\in\mathcal{N}\}} \sum_{c'=1}^{n_c^p} x_{t-1}^{[c',p]} \right),\end{aligned}$$

where $\mathcal{N}$, $\mathcal{S}$ denote the set of North and South pens, respectively, and n_c^p is the total number of individuals in pen $p = 1, 2, \ldots, P$. The function $\mathbb{1}_A$ is the indicator of event A.

We call the model described above the full model, denoted by $\mathcal{M}_1$. It may be of interest to treat the α values for North and South pens as common, and to test only for a distinct β value, or vice versa. This is meaningful as the value of α may relate to the location of the pen, whereas the value of β may be more closely linked to the area of the pen. Therefore, we compare the full model with three alternative models in order to establish which of the parameters α_s and α_n or β_s and β_n are common. Given the epidemic data, we aim to choose one from the following four candidate models:

Model $\mathcal{M}_1$: the full model with four parameters, α_s, α_n, β_s, and β_n.

Model $\mathcal{M}_2$: a special case of the full model, where the external transmission parameter is identical in the two sets of pens, i.e. $\alpha_s = \alpha_n = \alpha$.

Model $\mathcal{M}_3$: a special case of the full model, where the within-pen transmission parameter is identical in the two sets of pens, $\beta_s = \beta_n = \beta$.

Model $\mathcal{M}_4$: a special case of the full model with one parameter α, such that $\alpha_s = \alpha_n = \alpha$, and one parameter β, such that $\beta_s = \beta_n = \beta$.

5.3.2 Implementation details

In addition to the IS algorithm, we apply RJMCMC for comparison. The mechanism with which our proposed RJMCMC algorithm moves between models is illustrated in Figure 5.5. The prior in the model indicator $k \in \mathcal{K} = \{1, 2, 3, 4\}$ is the uniform distribution over the space of competing models, resulting in $\pi(k) = 0.25$ for all $k \in \mathcal{K}$.

The RJMCMC algorithm proceeds as follows. For a move from model $\mathcal{M}_1$ to $\mathcal{M}_2$ we keep parameters $\beta_s, \beta_n, m, \nu, \theta_R, \theta_F$ and $\mathbf{X}$ fixed and propose $\alpha_2 = \frac{L_n\,\alpha_{n_1} + L_s\,\alpha_{s_1}}{P}$, where L_n is the number of North pens, L_s is the number of South pens and P is the total number of pens. The Jacobian of this transformation is $\frac{L_n\,L_s}{P}$. For a reverse move, we need to increase the dimension of the parameter vector and so introduce an auxiliary random variable $u \sim \mathcal{N}(0, \sigma_1^2)$ with σ_1^2 fixed but chosen such that the algorithm can efficiently move between $\mathcal{M}_1$ and $\mathcal{M}_2$. Then, we set $\alpha_{s_1} = \alpha_2 + \frac{u}{L_s}$ and $\alpha_{n_1} = \alpha_2 - \frac{u}{L_n}$, and the Jacobian is $\frac{P}{L_n\,L_s}$. The acceptance probability of a move from model $\mathcal{M}_1$ to $\mathcal{M}_2$ is given by $\min(1, A_{12})$,

$$A_{12} = \frac{\pi(k=2, \boldsymbol{\psi}_2 \mid \mathbf{Y})}{\pi(k=1, \boldsymbol{\psi_1} \mid \mathbf{Y})}\left\{\frac{1}{\sigma_1\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma_1^2}\left(\frac{L_n\, L_s\,(\alpha_{s_1}-\alpha_{n_1})}{P}\right)^2\right]\right\}\frac{L_n\, L_s}{P},$$

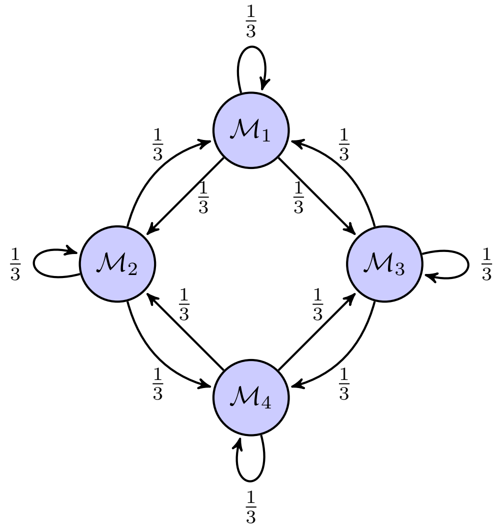
where $\boldsymbol{\psi}_1 = (\alpha_{s_1}, \alpha_{n_1}, \beta_s, \beta_n, m, \nu, \theta_R, \theta_F, \mathbf{X})$, $\boldsymbol{\psi}_2 = (\alpha_2, \beta_s, \beta_n, m, \nu, \theta_R, \theta_F, \mathbf{X})$, and $\pi(k, \boldsymbol{\psi}_k \mid \mathbf{Y})$ is the posterior probability of model $\mathcal{M}_k$, $k = 1, 2$. For the reverse move $\mathcal{M}_2$ to $\mathcal{M}_1$, the probability of accepting the jump is given by $\min(1, A_{21})$,

$$A_{21} = \frac{\pi(k=1, \boldsymbol{\psi}_1 \mid \mathbf{Y})}{\pi(k=2, \boldsymbol{\psi_2} \mid \mathbf{Y})}\left\{\frac{1}{\sigma_1\sqrt{2\pi}}\exp\left[-\frac{u^2}{2\sigma_1^2}\right]\right\}^{-1}\frac{P}{L_n\, L_s}.$$

Similar proposals are used for moves between models $\mathcal{M}_1$ and $\mathcal{M}_3$, $\mathcal{M}_2$ and $\mathcal{M}_4$ and finally $\mathcal{M}_3$ and $\mathcal{M}_4$. In all cases we choose the variance of the auxiliary variables u such that the probability of accepting a trans-dimensional move is at least 10%. In addition to the model-switching step, the within-model parameters are updated using either Gibbs or Hamiltonian Monte Carlo updates, similar to the algorithm used in the previous Section 5.2 for the Negative Binomial model. The algorithm runs for 400,000 iterations and we discard 50,000 as burn-in.

For IS, the proposal density $q(\boldsymbol{\theta}_k)$ is a mixture of the prior (5% mixing weight) and a multivariate Normal (95% mixing weight) with mean the estimated posterior mean vector and covariance the estimated covariance matrix from the

Figure 5.5: RJMCMC probabilities of moving from a model k to some model k', where $k, k' \in \mathcal{K} = \{1, 2, 3, 4\}$. The 4 competing models are used to investigate the effect of heterogeneity in colonisation rates. The models are described in Section 5.3.1.



MCMC output under model $\mathcal{M}_k,\ k = 1, 2, 3, 4$. MCMC samples are obtained as in within-model parameter updates of RJMCMC. Simulation studies regarding the performance of this algorithm in estimating the parameters of each model can be found in Section D.2.1 of the Appendix. For the proposal density $r(\mathbf{X}_k \mid \cdot)$ we use the FFBS algorithm to sample from the approximate full conditional $\pi(\mathbf{X}_k \mid \boldsymbol{\theta}_k, \mathbf{Y})$ (as discussed in Section 5.2.2) which can be written as:

$$\pi(\mathbf{X}_k \mid \boldsymbol{\theta}_k, \mathbf{Y}) = \prod_{p=1}^{P} \pi(\mathbf{X}_k^p \mid \boldsymbol{\theta}_k^p, \mathbf{Y}^p),$$

where $\mathbf{X}_k^p$ are the colonisation states of pen p and $\boldsymbol{\theta}_k^p$ the corresponding parameters of pen p under model $\mathcal{M}_k$. The construction of the importance sampling density for each one of the competing models is based on 20,000 MCMC samples after a burn-in of 5,000. For estimating the marginal likelihood we use 100,000 samples which requires roughly the same computational effort as RJMCMC.

Since the interpretation of m, κ, ν, θ_R and θ_F is common to all models, we use the same prior distributions for consistency with our prior beliefs; $m - 1, \kappa \sim \text{Ga}(0.01, 0.01)$ and $\nu, \theta_R, \theta_F \sim \text{Beta}(1, 1)$. Also, to each parameter $\alpha, \beta, \alpha_s, \alpha_n, \beta_s$ and β_n we assign a $\text{Ga}(1, 100)$ prior when necessary, because our simulation studies suggest that this prior results in accurate estimates of evidence (see Section D.2.2 of Appendix for more details). To obtain estimates of the Monte Carlo variability of the posterior model probabilities, we apply IS and RJMCMC 40 times.

5.3.3 Results

In Figure 5.6, we present the joint and marginal posterior distribution of the colonisation parameters. Generally, we observe considerable overlap in the α direction but separation in the β direction. Moreover, if model $\mathcal{M}_3$ ($\beta_s = \beta_n = \beta$) is used instead of model $\mathcal{M}_1$ ($\beta_s \neq \beta_n$), we obtain different results for parameters α_s and α_n, the main difference being that the posterior median of α_n is higher compared to the posterior median of α_s, as opposed to $\mathcal{M}_1$ where the posterior median $\alpha_n < \alpha_s$. This is in line with what would be expected, since now the parameters α_n and α_s of model $\mathcal{M}_3$ must take values that account for the number of infections occurring on the dataset, for which the common β cannot account. Specifically, under model $\mathcal{M}_3$ the parameters β_s and β_n are equal and therefore larger values of α_n explain the fact that there are more infections in North pens than in South pens and vice versa lower values of α_s account for the fact that there are fewer infection in the South pens. Nevertheless, there is considerable overlap of the distribution of α_n and α_s in both $\mathcal{M}_1$ and $\mathcal{M}_3$.

In contrast, when we allow parameters β_s and β_n to differ (under models $\mathcal{M}_1$ and $\mathcal{M}_2$), the within-pen transmission posterior median is found to be 2.5 times larger in North pens compared to South pens and there is only little overlap in the two distributions. Thus, there is evidence of differences between the North and South pens in terms of their within-pen transmission parameter but not in terms of their external transmission rate. However, to test formally whether the data provide evidence of heterogeneity in colonisation rates among South and North pens, we can use IS and RJMCMC to compute the Bayes factors, as described in Section 5.3.2.

Figure 5.7 shows the estimated posterior model probabilities for the two methods that we use. The RJMCMC and IS approaches give similar results and both support model $\mathcal{M}_2$ that has a median (over 40 implementations) posterior probability 0.7751 and 0.7673, respectively. Model $\mathcal{M}_4$ is the second most preferable and is assigned median posterior probabilities of roughly 0.16 by the two methods. Note that both models $\mathcal{M}_2$ and $\mathcal{M}_4$ allow for common external transmission rates between North and South pens. We also report the Bayes factors for the pairwise comparisons:

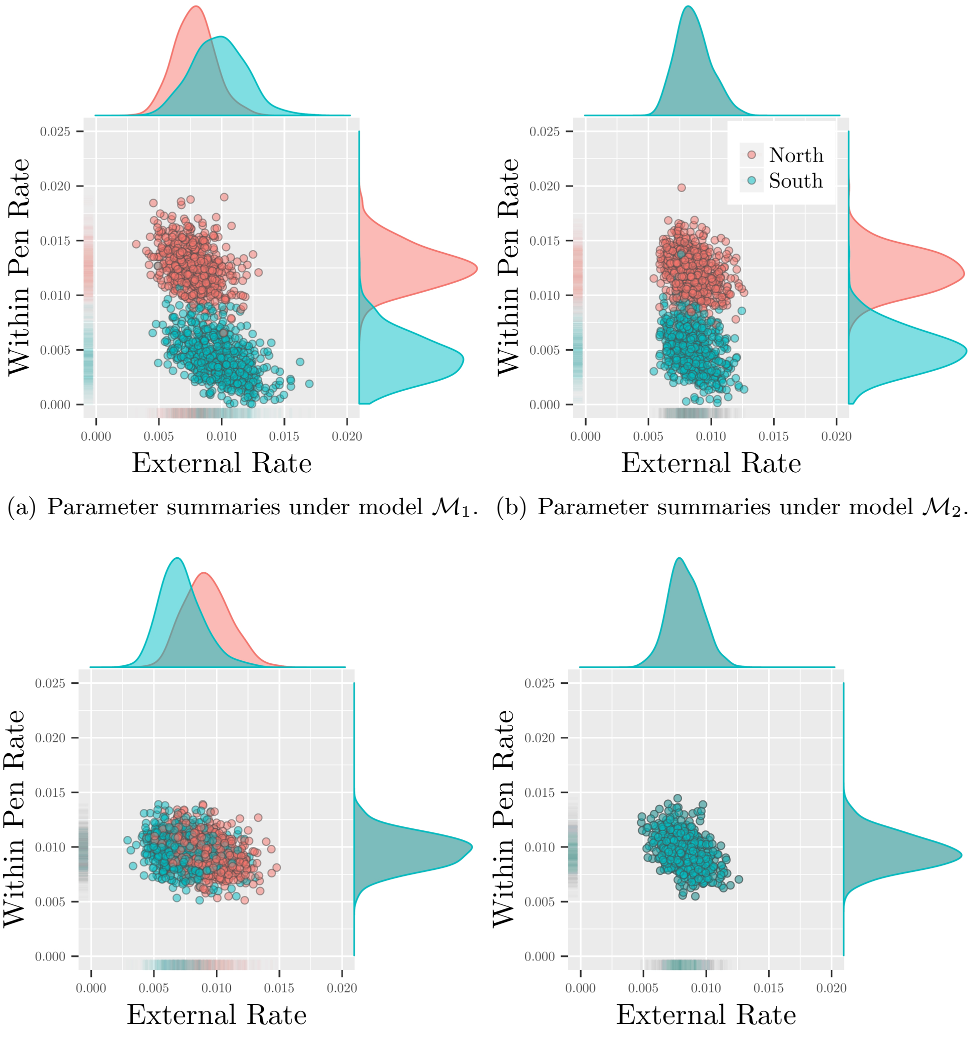
$$\begin{aligned} B_{21} &= 12.50, & B_{41} &= \ \ 2.61, \\ B_{23} &= 69.13, & B_{43} &= 14.43, \\ B_{24} &= \ \ 4.79, & B_{13} &= \ \ 5.53. \end{aligned}$$

The results give positive evidence in favour of model $\mathcal{M}_2$ when compared with models $\mathcal{M}_1, \mathcal{M}_4$, and strong evidence in favour of model $\mathcal{M}_2$ when compared with model $\mathcal{M}_3$. Overall, our analyses suggest that the data favour the hypothesis of equal external transmission rates between North and South pens but suggest that the within-pen transmission rate is higher in the North pen set compared to the South pen set.

5.3.4 Discussion

The IS and RJMCMC methods for computing Bayes factors are applied to investigate the effect of heterogeneity in colonisation rates among North and South pens. The two methods provide identical conclusions. Results favour model $\mathcal{M}_2$ with different within-pen transmission rates but common external colonisation rate between North and South pens. One explanation for that could be that the area is bigger in the South pens and so the animals make fewer contacts with each other or fewer contacts with infectious sources (e.g. faeces from colonised individuals) on

Figure 5.6: Joint and marginal posterior densities of the external and within-pen transmission rates, for *E. coli* dataset 1. The top left, top right, bottom left and bottom right panels represent models $\mathcal{M}_1 - \mathcal{M}_4$, respectively.



(a) Parameter summaries under model $\mathcal{M}_1$.

(b) Parameter summaries under model $\mathcal{M}_2$.

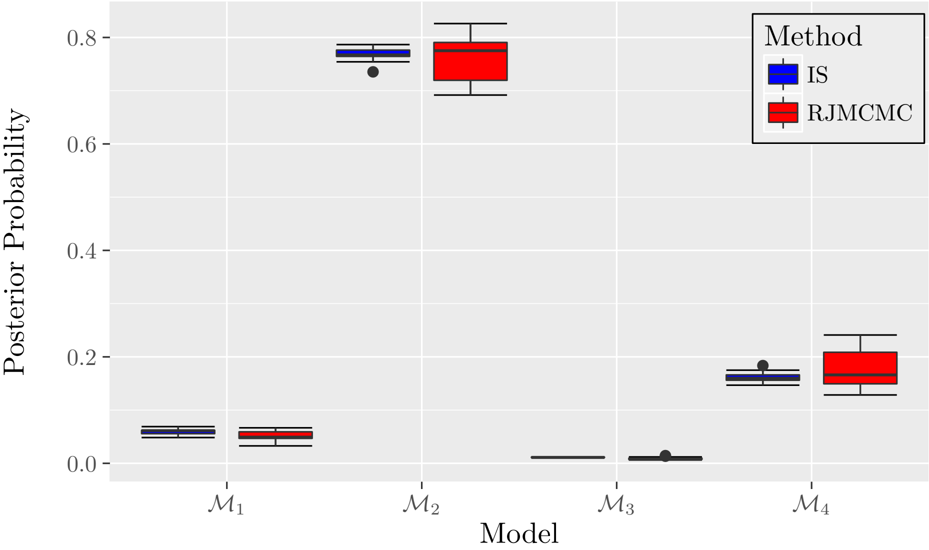
(c) Parameter summaries under model $\mathcal{M}_3$.

(d) Parameter summaries under model $\mathcal{M}_4$.

the ground.

The second most favourable model is $\mathcal{M}_4$, also allowing for common external transmission rate but with common within-pen transmission rates. These facts suggest that there is strong evidence for equal α values of North and South pens, meaning that pens with different geographical location are subject to the same levels of risk of *E. coli* O157:H7 infection from external sources. We can attribute this result to the large population (over 10,000) of cattle in the area around the facility in which our study was conducted. However, it should be emphasised that the design

Figure 5.7: Posterior probabilities of models $\mathcal{M}_1 - \mathcal{M}_4$ described in Section 5.3.1 used to analysed *E. coli* dataset 1, using IS and RJMCMC methods.



of the experiment was not set up to investigate different pen areas or locations and thus, the area in the pen is confounded with the location.

5.4 Investigating transmission between neighbouring pens

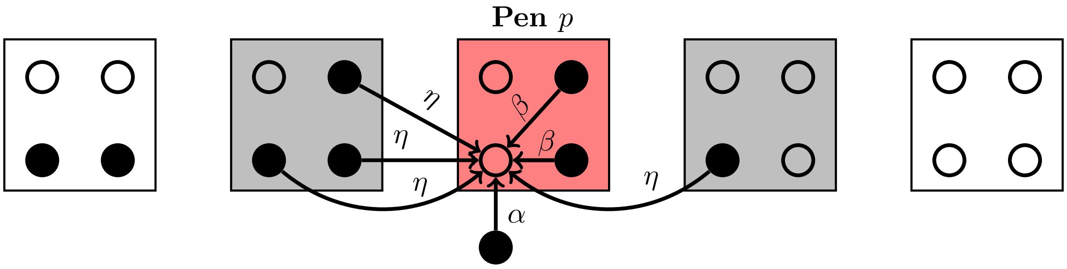
In this section, we consider a further model which allows for an extra source of colonisation namely transmission between neighbouring pens. Thus, we assume that the epidemic process mixes at an additional level, where each individual makes infectious contacts not only within a pen but also in its neighbourhood (pens that are located close by), with different colonisation rates in each setting. As before, we assume that individuals avoid infection from the community as a whole with a common probability.

This extension give us a new model with associated probability of a susceptible individual c housed in pen p being colonised at day t given by:

$$\mathbb{P}(X_t^{[c,p]} = 1 \mid X_{t-1}^{[c,p]} = 0) = 1 - \exp\Big(- \alpha - \beta \sum_{c'=1}^{n_c^p} x_{t-1}^{[c',p]} - \eta \sum_{\ell \in n(p)} \sum_{c'=1}^{n_c^p} x_{t-1}^{[c',\ell]} \Big),$$

where $n(p)$ is the set of neighbouring pens of pen p. Therefore, individuals living in different pens are not independent under this model. Figure 5.8 is a graphical representation of the potential transmission routes between colonised and susceptible individuals. We adopt this approach in light of our initial findings in Chapter 2, where for dataset 2 we observed a non-zero between-pen correlation that was higher for pens that are closer one with another. In Sections 5.4.2.1 and 5.4.2.2, we define the competing models considered for each dataset respectively. In the first dataset we assume a Negative Binomial colonisation period and different within-pen transmission rates for North and South pens, since we have found evidence in support of these assumptions. For dataset 2, we use a Geometric distribution since we have a sparse sampling interval.

Figure 5.8: Dynamics of infection in a population comprising of neighbouring (grey squares) and non-neighbouring pens (white squares) of pen p (red square). The state of an epidemic is shown for a given day, with solid circles indicating infected hosts and empty circles indicating susceptible hosts. The arrows represent potential transmission routes between infected and a given susceptible individual.



5.4.1 Implementation details

Inferences for this epidemic model are not straightforward because the underling colonisation processes are dependent across neighbouring pens. We fit the epidemic model using MCMC and data augmentation techniques that use similar ideas as in Section 5.2.2. Sampling directly from the full conditional distribution of the unobserved colonisation states in this case is complicated due to both: 1) FFBS being computationally infeasible to account for all possible interactions and 2) assuming a Negative Binomial colonisation period (for dataset 1). Therefore, we consider an independence sampler to simulate $\mathbf{X}$ with proposal given as the full conditional (π_H) under the Geometric model ($\kappa = 1$, to solve issue 2 when necessary) and with no interaction between pens ($\eta = 0$, to solve issue 1). To correct for the fact that we are using an approximate full conditional, we apply a Metropolis-Hastings correction. When η is included in a model, we update it using HMC, jointly with the other parameters for which HMC is employed.

We use the following prior distributions. For dataset 1 we set $\alpha, \beta_s, \beta_n \sim \text{Ga}(1,1)$, $m-1, \kappa \sim \text{Ga}(0.01, 0.01)$ and $\nu, \theta_R, \theta_F \sim \text{Beta}(1,1)$. For dataset 2 we have $\alpha, \beta \sim \text{Ga}(1,1)$, $m-1 \sim \text{Ga}(0.01, 0.01)$ and $\nu \sim \text{Beta}(1,1)$. We also set $\eta \sim$ Ga(1,1000), when η is included in a model. This prior is used in light of our simulations in Section D.3 of the Appendix where we found that this prior can successfully recover the true model. MCMC runs for 25,000 iterations and we save the last 20,000 based on which we construct the IS density of the model parameters $q(\cdot)$. A total of 150,000 samples are taken from this density. For $r(\cdot \mid \cdot)$ we use the same proposal as with MCMC that is, we take samples from the approximate distribution π_H, essentially assuming that $\eta = 0$ and $\kappa = 1$ (when necessary). We repeat IS 40 times to assess the variability of the estimator. RJMCMC is not applied for this comparison because it was hard to design an efficient proposal for the method to jump between the candidate models.

5.4.2 Results

5.4.2.1 Analysis of dataset 1

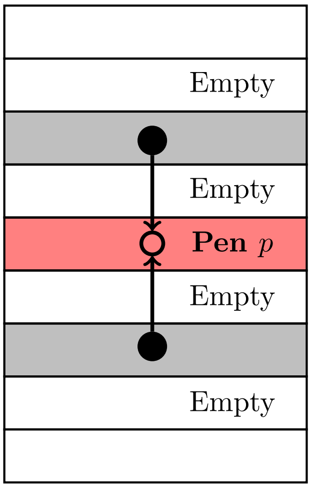
For this dataset we define as neighbours two pens that are only separated by an empty pen, as shown in Figure 5.9. Given the epidemic data, we aim to choose one from the following two models:

Model $\mathcal{M}_1$: the Negative Binomial model in which we account for two putative types of transmission, with common external colonisation rates α and within-pen colonisation rates β_n and β_s for North and South pens.

Model $\mathcal{M}_2$: the extended Negative Binomial model in which we account for three putative types of transmission, with external colonisation rate α, within-pen colonisation rates β_n, β_s and neighbour transmission rate η.

We report posterior summaries before we move to model selection. Some of the model parameter estimates can be found in Figure 5.10. The posterior densities of β_n, β_s, κ and m are very similar in both models. When model $\mathcal{M}_2$ is fit to the data, we observe a slight decrease in the value of α because η accounts for some of the external transmissions that under model $\mathcal{M}_1$ were captured by α. Nevertheless, the posterior median of η is 4×10^{-4}, which implies an average time between neighbouring pen infections of 2500 days. A formal model selection procedure provides positive support in favour of model $\mathcal{M}_1$ which has median posterior probability of 0.843 (the 95% CI over 40 replicates is [0.830, 0.858]), implying that there is no significant transmission of the disease between neighbours. The plot of the log

Figure 5.9: Grey squares denote pens that neighbour the red pen for *E. coli* O157:H7 dataset 1. The arrows represent potential transmission routes between infected individuals (solid circles) and neighbouring susceptible individual (empty circle).



marginal likelihoods of each model over iterations is given in Figure 5.11. Therefore, we have evidence that the experimental design that aimed to allow for no interaction between pens by separating them with an empty pen was successful.

5.4.2.2 Analysis of dataset 2

Once again the purpose of the analysis performed is to provide a model that better describes the transmission of *E. coli* O157:H7 infection in cattle. For this dataset, the layout of the pens is shown in Figure 2.2. There are several ways in which neighbours can be defined. We consider the following 5 definitions, each representing one model:

1. Pens that share only waterers (see Figure 5.12(a)).

2. Pens that share a boundary (waterer and/or wall) but not a feed bunk (see Figure 5.12(b)).

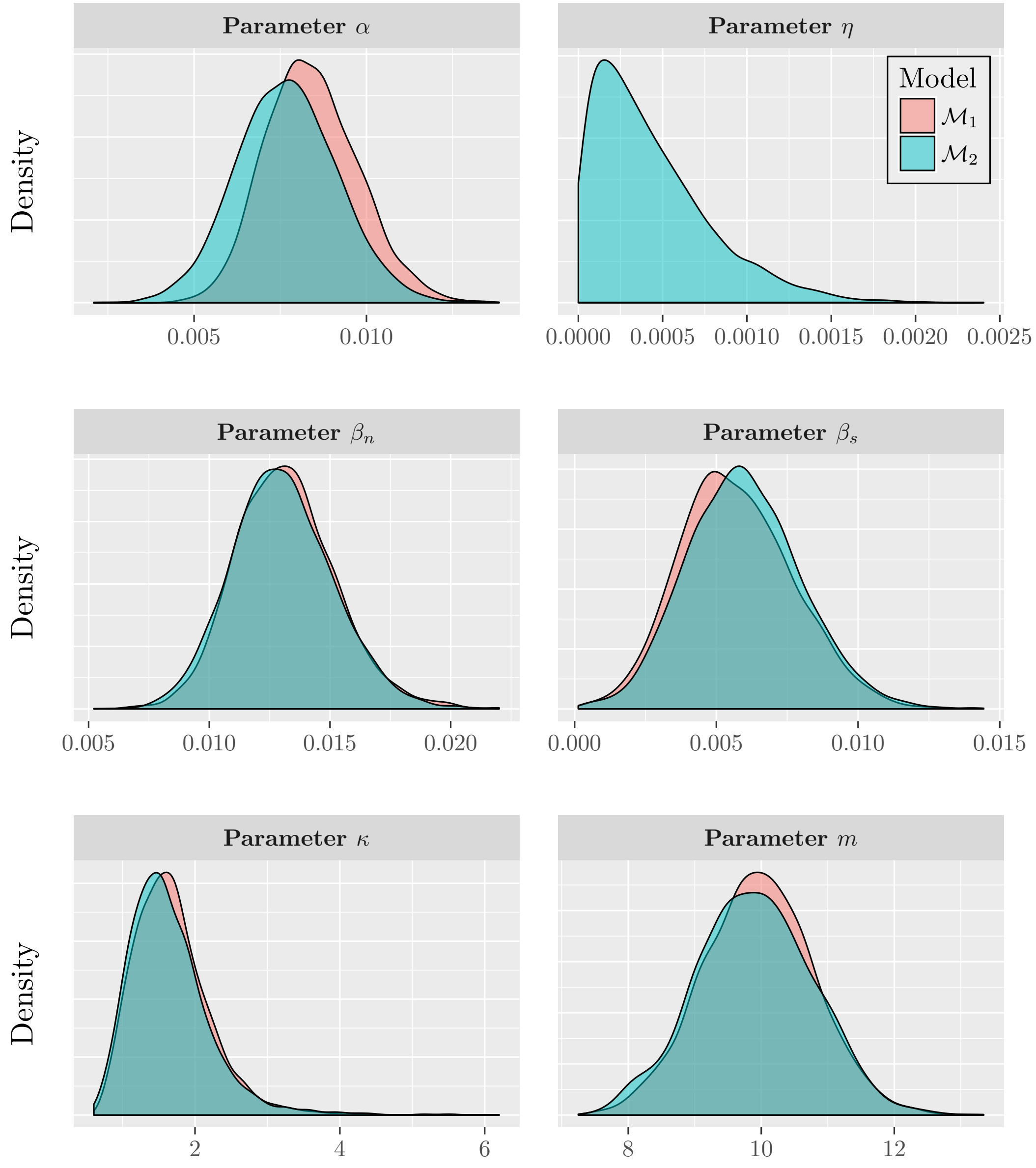
3. Pens that share a boundary (waterer and/or wall) and a feed bunk (see Figure 5.12(c)).

4. Pens that share only a feed bunk (see Figure 5.12(d)).

5. Pens that share a feed bunk and waterer (see Figure 5.12(e)).

In addition to the above models $\mathcal{M}_k$, where $k \in \{1, 2, 3, 4, 5\}$, we consider the basic model $\mathcal{M}_6$ which does not allow interactions between any pens ($\eta = 0$).

Figure 5.10: Marginal posterior densities of the main characteristics of the transmission of *E. coli* O157:H7 dataset 1, using model $\mathcal{M}_1$ and $\mathcal{M}_2$.



For all $\mathcal{M}_k$, $k = 1, \ldots, 6$, we use a Geometric distribution for the colonisation period because the sampling interval in dataset 2 is sparse (once every 2 weeks), and this makes inferences for the dispersion parameter in the Negative Binomial distribution challenging.

As discussed in Section 3.6.2, we fix the parameters θ_R and θ_F to 0.729 and

Figure 5.11: Log marginal likelihoods of models $\mathcal{M}_1$ (left panel) and $\mathcal{M}_2$ (right panel) of *E. coli* O157:H7 dataset 1, as obtained by the IS method. The shaded area corresponds to the 95% credible interval calculated over 40 replicates.

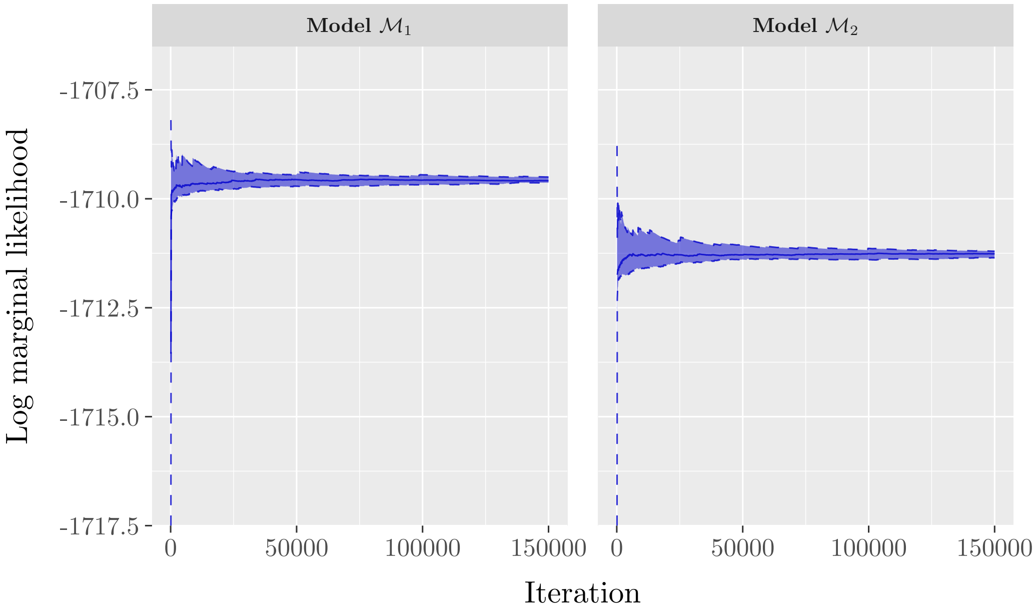
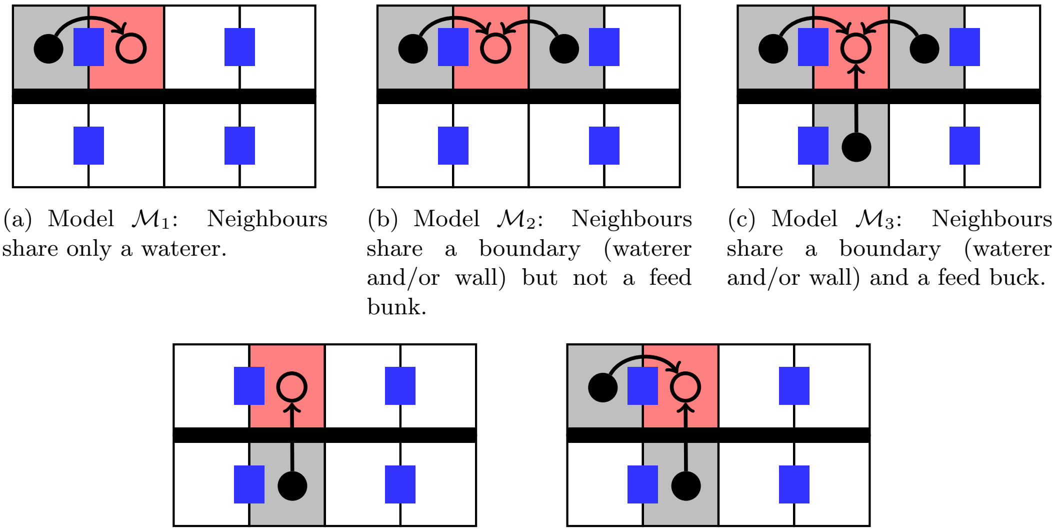


Figure 5.12: Grey squares denote pens that neighbour the red pen for *E. coli* O157:H7 dataset 2. The arrows represent potential transmission routes between infected individuals (solid circles) and neighbouring susceptible individual (empty circle).



(a) Model $\mathcal{M}_1$: Neighbours share only a waterer.

(b) Model $\mathcal{M}_2$: Neighbours share a boundary (waterer and/or wall) but not a feed bunk.

(c) Model $\mathcal{M}_3$: Neighbours share a boundary (waterer and/or wall) and a feed buck.

(d) Model $\mathcal{M}_4$: Neighbours share only a feed bunk.

(e) Model $\mathcal{M}_5$: Neighbours share a waterer and feed buck.

0.686, respectively. For the remaining model parameters and for each of the 6 models considered, we report posterior summaries in Table 5.5. Generally, the values depend on the model that we choose but there is considerable overlap of the 95% credible intervals among the 6 competing models. The most notable differences are found for the external transmission rate α and the between neighbour transmission rate η. Notably, the largest value of η is obtained under $\mathcal{M}_1$. However, no definitive conclusions can be reached using these summaries and hence a formal model comparison is required to measure the evidence in support of each model.

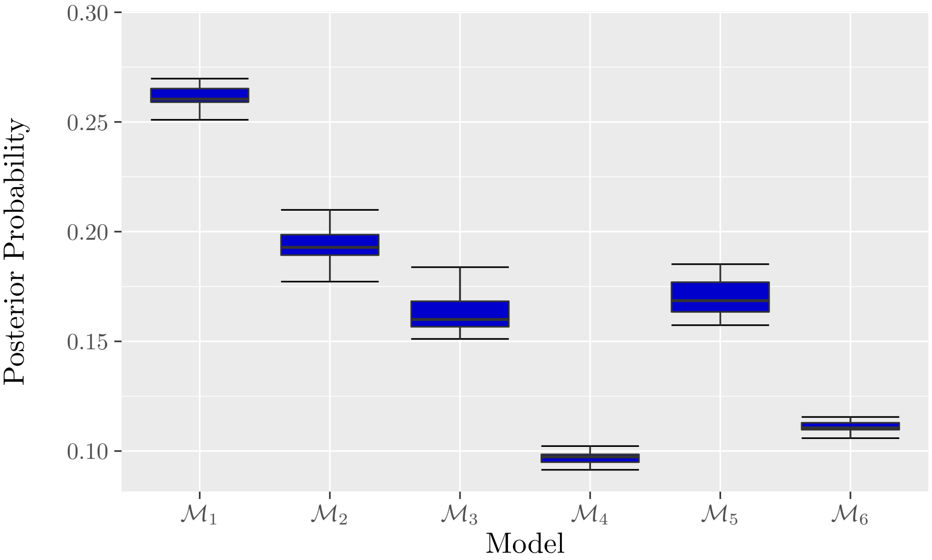
Table 5.5: Parameter estimation for *E. coli* O157:H7 dataset 2. Posterior mean, standard deviation, median and 95% credible interval for the parameters of each model.

Model	Mean	s.d.	2.5%	Median	97.5%
Model $\mathcal{M}_1$					
α	0.00354	0.00105	0.00198	0.00343	0.00572
β	0.01051	0.00258	0.00688	0.01015	0.01723
δ	0.00177	0.00136	0.00022	0.00159	0.00409
m	14.0225	2.42279	6.93560	14.1664	18.2691
ν	0.05108	0.01760	0.02260	0.04901	0.09109
Model $\mathcal{M}_2$					
α	0.00332	0.00114	0.00164	0.00321	0.00567
β	0.01058	0.00263	0.00705	0.01020	0.01736
δ	0.00097	0.00076	0.00010	0.00087	0.00217
m	14.2041	2.37620	6.98463	14.3635	18.2945
ν	0.05094	0.01745	0.02240	0.04908	0.09023
Model $\mathcal{M}_3$					
α	0.00312	0.00117	0.00139	0.00302	0.00543
β	0.01077	0.00273	0.00721	0.01035	0.01791
δ	0.00067	0.00054	0.00007	0.00061	0.00149
m	14.1204	2.47629	6.81213	14.2840	18.3770
ν	0.05157	0.01784	0.02266	0.04945	0.09273

Model $\mathcal{M}_4$					
α	0.00374	0.00105	0.00221	0.00363	0.00594
β	0.01107	0.00270	0.00758	0.01062	0.01856
δ	0.00086	0.00105	0.00004	0.00069	0.00223
m	14.2945	2.46251	6.96882	14.4856	18.5131
ν	0.05123	0.01762	0.02274	0.04927	0.09155
Model $\mathcal{M}_5$					
α	0.00338	0.00108	0.00179	0.00327	0.00563
β	0.01077	0.00265	0.00719	0.01039	0.01765
δ	0.00082	0.00071	0.00008	0.00073	0.00189
m	14.0675	2.42564	6.87511	14.2436	18.2196
ν	0.05136	0.01772	0.02248	0.04924	0.09182
Model $\mathcal{M}_6$					
α	0.00430	0.00102	0.00279	0.00420	0.00632
β	0.01127	0.00293	0.00781	0.01081	0.02027
m	14.4317	2.36214	7.36671	14.5578	18.5148
ν	0.05081	0.01757	0.02215	0.04888	0.09042

The result of the model comparison is shown in Figure 5.13. Model $\mathcal{M}_1$ has the highest median (over 40 replicates) posterior probability 0.263, followed by model $\mathcal{M}_2$ with 0.194, $\mathcal{M}_5$ with 0.170, $\mathcal{M}_3$ with 0.160, $\mathcal{M}_6$ with 0.113 and $\mathcal{M}_4$ with posterior probability 0.100. Thus, within this set of candidates epidemiological models, the most suitable model to describe the data is the one where neighbours share a waterer and no other boundary or feed bunk. Further, models $\mathcal{M}_2$, $\mathcal{M}_5$ and $\mathcal{M}_3$ that are the second, third and fourth most favourable by the algorithm, also assume that neighbours share their waterers. Therefore, there is evidence in favour of the hypothesis that the transmission is enhanced when two pens share a waterer, which confirms that contaminated waterers are a potential route of *E. coli* O157:H7 transmission (Faith *et al.*, 1996; Rahn *et al.*, 1997; Shere *et al.*, 2002). Interestingly, model $\mathcal{M}_4$ was the least preferable model, even compared to the simple model $\mathcal{M}_6$. A possible explanation for this result is that the design of the study did not allow for any contact between animals through the feed bunk.

Figure 5.13: Posterior probabilities of models $\mathcal{M}_1 - \mathcal{M}_6$ described in Section 5.4.2.2 used to analysed *E. coli* dataset 2, using IS method.



5.4.3 Discussion

The objective of this section was to identify possible routes of *E. coli* O157:H7 transmission that relate to the arrangement of pens near one another. For the first dataset, this question connects to the experimental design of the study which aimed to prevent transmission between pens. To achieve so, neighbouring pens were separated by an empty pen ensuring that each had its own water supply and feed bunk. We found that this design was successful since our model comparison showed positive evidence that transmission of the disease was not possible between pens that were nearby.

For the second dataset, by contrast, pens were located next to one another and some of them shared waterers and feed bunks in pairs. In this case, we have found positive evidence that *E. coli* O157:H7 is more likely to be transmitted between pens that share a waterer than to those that share a boundary and/or a feed bunk. This result confirms previous findings which indicate that contaminated drinking water may contribute to the spread of the disease (Faith *et al.*, 1996; Rahn *et al.*, 1997; Shere *et al.*, 2002).

5.5 Discussion

Since much is unknown about the exact mechanism of *E. coli* O157:H7 transmission, the first stage of our analysis is to determine an appropriate model using model selection. Therefore, in this chapter we developed a set of stochastic epidemic models that represent different assumptions regarding the transmission of infection among individuals and we compared them using Bayes factors. The RJMCMC and IS methods were implemented to estimate posterior model probabilities and marginal likelihoods, respectively, for different sets of competing models in the light of two observed datasets and simulated datasets (see also Appendix D).

Via simulations, we illustrated that RJMCMC and IS methods perform very well in providing an accurate estimate of the evidence when a suitable prior is used. However, it is challenging to design efficient jump mechanisms for RJMCMC algorithms when we consider four or more competing models, and therefore IS was found to be a useful alternative to estimate Bayes factors. Even though IS requires only within-model MCMC, it is still sometimes challenging to determine good proposal densities, especially for the unobserved data. In our examples we have shown how to design efficient proposals for the hidden states, either directly from the full conditional or using an approximate full conditional distribution.

We explored the use of a Negative Binomial distribution for the colonisation period of cattle as an alternative to the traditionally employed Geometric model. A formal statistical comparison between the two models favoured the Negative Binomial with weak evidence. This implies the probability that an individual clears the disease increases, the longer it remains colonised. The biological interpretation of our finding might be that cattle develop an immune response to the bacteria of *E. coli* O157:H7.

In the above modelling framework, we assumed that all pens have the same risk of acquiring colonisation from outside the pen and transmitting infection within the pen. We assessed the effect of relaxing this assumption, by allowing the rates of external and within-pen transmission to differ between pens according to their size and location. The analysis demonstrated that there is real evidence of differences between pens in terms of their within-pen colonisation rates but not their risk of acquiring the disease from external sources. In particular, we estimated that the smaller North pens have a higher within-pen colonisation rate which might arise due to there being more contacts between individuals compared to large pens that have bigger area and hence less contacts. An extension of this work will consider a hierarchical model in which each pen is allowed to have its own within-pen and external rates, sampled from some hyper-distribution whose parameters would then be subject to inference.

Finally, motivated by some previous studies which hypothesise that animal feed and waterers in particular are important sources of *E. coli* O157 transmission (Faith *et al.*, 1996; Rahn *et al.*, 1997; Shere *et al.*, 2002; Cobbold and Desmarchelier, 2002), we extend our models to account for such incidents. Our results confirm the hypothesis that contaminated waterers contribute to the spread of the pathogen since we found appreciable evidence in support of this model.

Although the models that we consider in this chapter were designed to analyse the dynamics of *E. coli* in cattle, they can be adopted to model other kinds of population heterogeneity or diseases in other livestock or humans. For example, the methodology presented can be applied to investigate whether children are more prone to influenza infections through the family, community or school. In all of these cases, the techniques for model selection that we have presented are directly applicable and hence can be employed to address various epidemiological questions of interest as was illustrated in this chapter.

Chapter 6

Scalable Inference For Coupled Hidden Markov And Semi-Markov Epidemic Models

6.1 Introduction

Hidden Markov models are among the most widely used approaches for modelling time series data, when it can be assumed that the observed data are indicative of some underlying hidden state. In the basic HMM, a single variable represents the state of the system at any time. However, many interesting systems are composed of multiple interacting processes, and thus this single-process model is not appropriate. An example, and the main motivation for the work presented in this chapter, is the spread of infection at an individual level; individuals in a population can jump between several epidemic states (for instance infected or susceptible), and change state according some probability distribution that depends on the previous hidden state of all individuals, including themselves. Therefore, the epidemic model is composed of multiple Markov chains, one per individual, which influence and interact with each other.

Various extended HMM models have been proposed to solve coupled, multiple chain data analysis problems. These extensions basically factor the HMM state into a collection of state variables, $\mathbf{X}_t^{[1:C]} = \left(X_t^{[1]}, X_t^{[2]}, \ldots, X_t^{[C]}\right)$, where $X_t^{[c]} \in \mathcal{X}_s = \{s_1, s_2, \ldots, s_N\}$ corresponds to the hidden state of chain c at time t. In our case, we use coupled hidden Markov models (CHMMs) to capture these interactions (Brand, 1997), where the current state of a chain depends on the previous state of all the chains. The temporal dependence between the hidden states is captured though matrices of conditional probabilities which couple the chains.

This structure implies that the state space grows exponentially with respect to the number of chains and thus exact inference quickly becomes computationally intractable. Thus, with more and more data becoming available it is extremely important to design CHMM algorithms that scale well for big datasets. As a consequence, inference in CHMM has received much attention and a variety of techniques are available. Existing methods for inference are either based on various likelihood approximation schemes or achieved by MCMC sampling methods.

The inference problem for CHMMs usually includes both hidden state and parameter estimation. Early literature on the topic was focused on maximum likelihood estimation, achieved using an EM algorithm. Among this class, many researches proposed several variations of the standard CHMM for which inference problems become more tractable. For instance, Brand *et al.* (1997) made a simplification that the transition probability for a given chain conditional on the others is separable and therefore can be represented as the product of all marginal conditional probabilities:

$$\mathbb{P}\left(X_t^{[c]} \mid \mathbf{X}_{t-1}^{[1:C]}\right) = \prod_{c'=1}^{C} \mathbb{P}\left(X_t^{[c]} \mid X_{t-1}^{[c']}\right).$$

One issue with this formulation is that the right hand side does not sum up to one. To overcome this issue, Saul and Jordan (1999) and Zhong and Ghosh (2002), made a different simplifying assumption that the transition probability is replaced with a weighted sum of the marginal conditional probabilities,

$$\mathbb{P}\left(X_t^{[c]} \mid \mathbf{X}_{t-1}^{[1:C]}\big) = \sum_{c'=1}^{C} \phi_{c'\,c}\,\mathbb{P}\big(X_t^{[c]} \mid X_{t-1}^{[c']}\right),$$

where $\phi_{c'\,c}$ is the coupling weight from chain c' to chain c. Another approach includes the work by Kwon and Murphy (2000), in which the authors used CHMMs to model freeway traffic and considered two different approximations for the E-step, one based on the particle filtering, and the other based on the Boyen-Koller algorithm.

All of these approaches have been developed assuming specialised CHMMs, in order to reduce the computational complexity; however, applications in Brand *et al.* (1997), Saul and Jordan (1999) and Zhong and Ghosh (2002) involved only two chains. Choi *et al.* (2013) recently proposed a fast algorithm for sparsely correlated HMMs where inference for each individual chain is performed conditioning on the hidden state vectors in all other chains, by assuming sparsity when incorporating correlations between the current chain and all the other chains. The approach presented in this study is similar to Choi *et al.* (2013), however in their work the authors used the standard forward backward algorithm to deterministically select the hidden state sequence in each chain. In contrast, we are interested in performing CHMM inference under a Bayesian paradigm.

The second class of methods consists of MCMC approaches. One considerable challenge among this class concerns the computation of the posterior distribution of the hidden states conditional on the observed data and model parameters, and many techniques have been proposed to overcome this challenge. The most popular approach to exact Monte Carlo inference can be achieved by converting the CHMM into an equivalent HMM with N^C states (where N is the total number of possible hidden states) and applying the standard forward filtering backward sampling algorithm (Carter and Kohn, 1994; Chib, 1996).

However, even though implementation of FFBS is quite efficient for HMMs with a moderately large number of states N, it can be computationally prohibitive for CHMMs with only a small number of chains C. As a result, several alternative methods have been to solve the problem including conditional single-site (Dong *et al.*, 2012) or block updates (specifically for epidemics, Spencer *et al.* 2013). While these methods are less computationally demanding than the FFBS, they typically produce highly correlated samples.

Assuming a sparse transition matrix is one way to speed up FFBS algorithm, and such a method was recently proposed by Sherlock *et al.* (2013). In this work, the authors impose a structure on each chain's transition matrix, with transition probabilities depending on covariates through logistic regression. These covariates include the states of the other chains and other external factors. While this approach reduces computation time, it requires the structure of transition matrices to be estimated or known in advance.

The computationally most demanding part of the FFBS algorithm is the summation over all possible configurations of the hidden state variables within each time step, which is done during the forward filtering step. The motivation behind this work is to avoid doing this calculation by proposing a slightly modified conditional forward variable which can be calculated for each chain and can reduce the computational complexity to a practical level. In particular, we propose a Gibbs sampling algorithm for the CHMM which is based on simulating from the posterior conditional distribution over a single chain given the rest.

The main contribution of this study is the development of a novel extension of the FFBS algorithm which explicitly takes into account the interaction between chains, without imposed any structure on the transition matrix, and at the same time achieving a balance between sampling efficiency and computational complexity. We initially restrict our presentation to CHMMs, and we subsequently describe how the proposed method can be extended for models with a richer structure in their set of hidden state variables. In particular, we show how our analysis can be applied to coupled hidden semi-Markov models (CHSMMs), where the hidden process persists in the same state for some non-Markov duration, and to models consisting of interacting coupled hidden Markov processes. This creates a novel class of algorithms that are computationally efficient and at the same time provide reliable results.

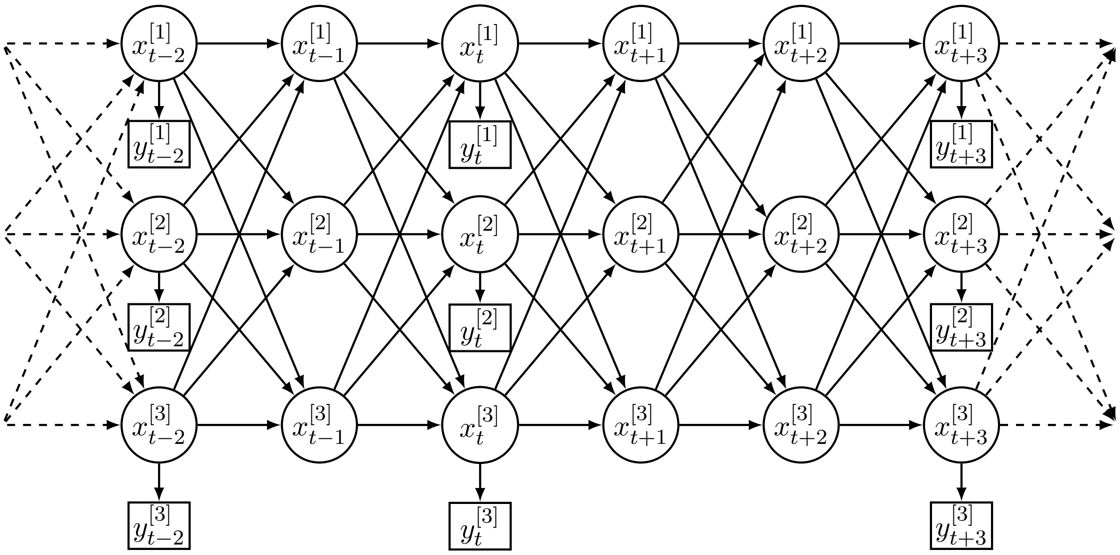
The remainder of this chapter is organised as follows. In Section 6.2 we describe more formally the coupling structure in our CHMM model. Sections 6.3 and 6.4 describe existing MCMC methods in the literature for estimating CHMMs, and summarise their main advantages and limitations. In this section two new algorithms for inference in CHMMs are proposed. In Section 6.5 we put CHMMs in the context of modelling the spread of infectious diseases. Section 6.6 illustrates the results using simulation studies with different degrees of complexity that compare the MCMC methods with respect to the mixing, efficiency and computational requirement. Results shows that our approach allows the FFBS algorithm to be used with much large populations than has previously been possible and is linear in the population size rather than exponential. The methodology is also used to analyse the real *E. coli* O157:H7 datasets, in Section 6.7. In Section 6.8 we conclude with some discussion and possible extensions.

6.2 Coupled hidden Markov model

A coupled hidden Markov model is a collection of many HMMs, which are coupled with some temporal dependency structure of the hidden states. There are two conditional independence assumptions made about the observations and states. As in HMMs, given the value of its hidden state, one observation is independent of all other states and observations in the CHMM. The difference with HMMs is that one hidden state is not only dependent on the previous state in this chain, but also on the previous state of all other chains. The latter dependence constitutes the interaction between the multiple chains.

We choose the coupling structure shown in Figure 6.1 in our CHMM. More formally, we use $X_t^{[c]}$ to denote the hidden state variable of chain $c \in \{1, 2, \ldots, C\}$ at time $t \in \{1, 2 \ldots, T\}$ with a finite set of possible states. For simplicity, we assume that all chains share the same set of possible states; nevertheless, the method can be easily extended to the more general case where chains do not share the same state space. Therefore, we assume without loss of generality that $X_t^{[c]} \in \mathcal{X}_s = \{s_1, s_2, \ldots, s_N\}, N \geq 1$. For example, the simplest model is a binary chain with $N = 2$, where the two states $s_1 = 0$ and $s_2 = 1$ correspond to susceptible and

Figure 6.1: A coupled hidden Markov model represented as a dynamic Bayesian network, with three hidden chains ($C = 3$) and possibly several missing observations. Circle nodes denote hidden states, square nodes denote observations, and the arrows between nodes reflect the probabilistic dependencies between random variables.



infected individuals. We consider non-homogeneous Markov chains in which the transition probabilities depend on time and thus for all $\ell, k \in \mathcal{X}_s$ we define:

$$\mathbb{P}\left(X_t^{[c]} = k \mid X_{t-1}^{[c]} = \ell,\, \mathbf{X}_{t-1}^{[-c]},\, \boldsymbol{\theta}\right) = P_{\ell,\,k,\,t}^{[c]} \tag{6.1}$$

where $\mathbf{X}_{t-1}^{[-c]}$ denotes $\left(X_{t-1}^{[1]}, X_{t-1}^{[2]}, \ldots, X_{t-1}^{[C]}\right)$ with $X_{t-1}^{[c]}$ removed. To fully define the distribution of the state, a marginal distribution for $X_1^{[c]}$ needs to be specified, $\mathbb{P}\left(X_1^{[c]} = k \mid \boldsymbol{\theta}\right) = \nu_k^{[c]}$, for $c = 1, 2, \ldots C$.

The state of each chain is not directly observable. All there is, as in HMMs, is an observation $Y_t^{[c]}$ associated with the unobserved state $X_t^{[c]}$. The relation between $X_t^{[c]}$ and $Y_t^{[c]}$ will differ depending on the application, but depending on the value of the state we can write:

$$\pi\left(Y_t^{[c]} = y_t^{[c]} \mid X_t^{[c]} = k,\, \boldsymbol{\theta}\right) = f_k\left(y_t^{[c]} \mid \boldsymbol{\theta}\right), \quad k \in \mathcal{X}_s \tag{6.2}$$

which can be discrete or continuous. If $y_t^{[c]}$ is empty due to missing data, we set $f_k\left(y_t^{[c]} \mid \boldsymbol{\theta}\right)$ to 1. The parameters of this model, $\boldsymbol{\theta}$, will be the parameters of the observed and hidden processes of Equations (6.2) and (6.1), respectively.

6.3 Bayesian analysis and MCMC methods

One considerable challenge on estimating CHMMs is that the likelihood function of the observed data given the model parameters is computationally intractable for even moderate numbers of states or interacting chains. One solution to this problem is to impute the hidden states using data augmentation MCMC. This approach yields a joint posterior density for the unobserved states and the model parameters that is known up to proportionality,

$$\begin{aligned}
&\pi\left(\boldsymbol{\theta}, \mathbf{X}_{1:T}^{[1:C]} \mid \mathbf{Y}_{1:T}^{[1:C]}\right) \\
&\quad \propto \pi(\boldsymbol{\theta})\, \mathbb{P}\left(\mathbf{X}_1^{[1:C]} \mid \boldsymbol{\theta}\right) \left(\prod_{t=2}^{T} \mathbb{P}\left(\mathbf{X}_t^{[1:C]} \mid \mathbf{X}_{t-1}^{[1:C]}, \boldsymbol{\theta}\right)\right) \left(\prod_{t=1}^{T} \pi\left(\mathbf{Y}_t^{[1:C]} \mid \mathbf{X}_t^{[1:C]}, \boldsymbol{\theta}\right)\right) \\
&\quad = \pi(\boldsymbol{\theta}) \left(\prod_{c=1}^{C} \mathbb{P}\left(X_1^{[c]} \mid \boldsymbol{\theta}\right)\right) \left(\prod_{t=2}^{T}\prod_{c=1}^{C} \mathbb{P}\left(X_t^{[c]} \mid X_{t-1}^{[c]}, \mathbf{X}_{t-1}^{[-c]}, \boldsymbol{\theta}\right)\right) \\
&\qquad \times \left(\prod_{t=1}^{T}\prod_{c=1}^{C} \pi\left(Y_t^{[c]} \mid X_t^{[c]}, \boldsymbol{\theta}\right)\right) \qquad (6.3)
\end{aligned}$$

by assuming that a prior for the parameters, $\pi(\boldsymbol{\theta})$, has been specified. In order to simplify the notation it is convenient to adopt the following conventions:

$$\begin{aligned}
\mathbf{X}_t^{[1:C]} &= \left(X_t^{[1]}, X_t^{[2]}, \ldots, X_t^{[C]}\right) \\
\mathbf{X}_{1:t}^{[1:C]} &= \left(\mathbf{X}_1^{[1:C]}, \mathbf{X}_2^{[1:C]}, \ldots, \mathbf{X}_t^{[1:C]}\right)
\end{aligned}$$

with similar conventions applying to $\mathbf{Y}_t^{[1:C]}$ and $\mathbf{Y}_{1:t}^{[1:C]}$. In words, $\mathbf{X}_t^{[1:C]}$ is the set of the hidden states of all chains within a time step t, and $\mathbf{X}_{1:t}^{[1:C]}$ denotes the whole hidden state process up to time t.

As discussed in Section 1.2.2.3, samples from the joint posterior of the model parameters and the hidden states are generated as follows: first, $\boldsymbol{\theta}$ is updated conditional on the current values of $\mathbf{X}_{1:T}^{[1:C]}$ and then, $\mathbf{X}_{1:T}^{[1:C]}$ is updated conditional on $\boldsymbol{\theta}$. In this chapter, we focus on the problem of updating the hidden states, which is the most computational demanding part.

6.4 Updating the hidden states

Before discussing the details of our new approaches in Sections 6.4.4 and 6.4.5, we first briefly describe the standard algorithms for the CHMMs within this framework.

For further details regarding these methods, we refer the readers to the original publications.

6.4.1 Single-site Gibbs updates

The simplest way to update the matrix of hidden states $\mathbf{X}_{1:T}^{[1:C]}$ is to draw each one of the $C \times T$ components from its full conditional distribution:

$$\begin{aligned}\mathbb{P}&\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{-t}^{[-c]} = \mathbf{x}_{-t}^{[-c]}, \mathbf{Y}_{1:T}^{[1:C]} = \mathbf{y}_{1:T}^{[1:C]}, \boldsymbol{\theta}\right) \\ &\propto \mathbb{P}\left(X_t^{[c]} = x_t^{[c]}, \mathbf{X}_{-t}^{[-c]} = \mathbf{x}_{-t}^{[-c]}, \mathbf{Y}_{1:T}^{[1:C]} = \mathbf{y}_{1:T}^{[1:C]} \mid \boldsymbol{\theta}\right), \qquad (6.4)\end{aligned}$$

for $x_t^{[c]} \in \mathcal{X}_s$, where the normalising constant is the sum of all the terms over $X_t^{[c]}$ from s_1 to s_N, and $\mathbf{X}_{-t}^{[-c]}$ denotes the whole hidden state process excluding $X_t^{[c]}$. Note that these terms are calculated as in Equation (6.3) without the prior on $\boldsymbol{\theta}$. This method is called single-site update and has been used by Dong *et al.* (2012) for modelling the spread of infection within a social network (see Figure 6.3(b), page 127).

Note that Equation (6.4) can be further simplified by using the conditional independence assumptions; however, we choose to keep the general form for use in later sections. In our case, we only need to examine the states of a few neighbouring node to sample from $X_t^{[c]}$, as shown in Figure 6.2; each node is conditionally independent of all other nodes given its Markov blanket. Therefore, the full conditional distribution for $X_t^{[c]}$, for $t = 2, 3, \ldots, T-1$, simplifies to:

$$\begin{aligned}\mathbb{P}&\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{-t}^{[-c]} = \mathbf{x}_{-t}^{[-c]}, \mathbf{Y}_{1:T}^{[1:C]}, \boldsymbol{\theta}\right) \\ &= \mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{t-1}^{[1:c]} = \mathbf{x}_{t-1}^{[1:c]},\, \mathbf{X}_t^{[-c]} = \mathbf{x}_t^{[-c]},\, \mathbf{X}_{t+1}^{[1:c]} = \mathbf{x}_{t+1}^{[1:c]},\, Y_t^{[c]}, \boldsymbol{\theta}\right) \\ &\propto \mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid X_{t-1}^{[c]} = x_{t-1}^{[c]},\, \mathbf{X}_{t-1}^{[-c]} = \mathbf{x}_{t-1}^{[-c]}, \boldsymbol{\theta}\right) \pi\left(Y_t^{[c]} \mid X_t^{[c]} = x_t^{[c]}, \boldsymbol{\theta}\right) \\ &\quad \times \prod_{c'=1}^{C} \mathbb{P}\left(X_{t+1}^{[c']} = x_{t+1}^{[c']} \mid X_t^{[c']} = x_t^{[c']},\, \mathbf{X}_t^{[-c']} = \mathbf{x}_t^{[-c']}, \boldsymbol{\theta}\right) \\ &= P^{[c]}_{x_{t-1}^{[c]}, x_t^{[c]}, t} \left[\prod_{c'=1}^{C} P^{[c']}_{x_t^{[c']}, x_{t+1}^{[c']}, t+1}\right] f_{x_t^{[c]}}\left(y_t^{[c]} \mid \boldsymbol{\theta}\right),\end{aligned}$$

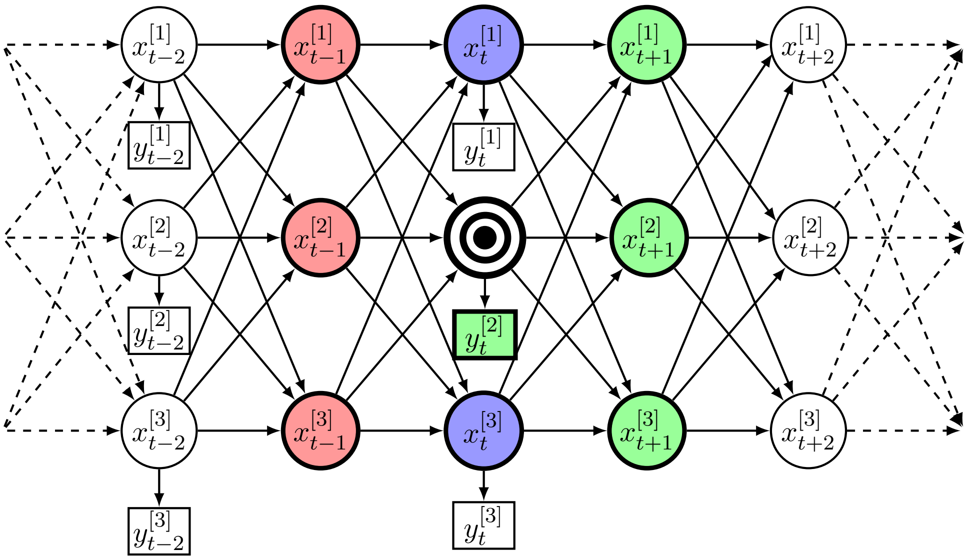
for $x_t^{[c]} \in \mathcal{X}_s$. Similarly, we obtain the full conditional distributions for $X_1^{[c]}$ and $X_T^{[c]}$. Therefore, we have that:

$$\mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{-t}^{[-c]} = \mathbf{x}_{-t}^{[-c]}, \mathbf{Y}_{1:T}^{[1:C]}, \boldsymbol{\theta}\right)$$

$$\propto \begin{cases} \nu^{[c]}_{x_t^{[c]}} \left[\prod_{c'=1}^{C} P^{[c']}_{x_t^{[c']}, x_{t+1}^{[c']}, t+1} \right] f_{x_t^{[c]}} \left(y_t^{[c]} \mid \boldsymbol{\theta} \right) & t = 1, \\ P^{[c]}_{x_{t-1}^{[c]}, x_t^{[c]}, t} \left[\prod_{c'=1}^{C} P^{[c']}_{x_t^{[c']}, x_{t+1}^{[c']}, t+1} \right] f_{x_t^{[c]}} \left(y_t^{[c]} \mid \boldsymbol{\theta} \right) & t = 2, 3, \ldots, T-1, \\ P^{[c]}_{x_{t-1}^{[c]}, x_t^{[c]}, t} \, f_{x_t^{[c]}} \left(y_t^{[c]} \mid \boldsymbol{\theta} \right) & t = T, \end{cases}$$

where the normalising constant in each case is the sum of all the terms over $X_t^{[c]}$ from s_1 to s_N.

Figure 6.2: The Markov blanket of a node (target) consists of all nodes that make this node conditionally independent of all the other nodes in the model: the parents (red nodes), the children (green nodes) and the parents of the children (blue nodes).



Thus we need to calculate $C \times T$ variables and each one requires $\mathcal{O}(N)$ time to compute giving an overall complexity of $\mathcal{O}(CNT)$. Despite being easy in implementation, it has been shown by Scott (2002) that the single-site update algorithm leads to extremely slow mixing in the resulting MCMC chains. This fact is due the high temporal dependence in the hidden state process.

6.4.2 Block proposals updates

Recently, Spencer *et al.* (2015) developed a method for epidemic models which proposes to change blocks of state components within a single chain, based on their current values. This method is a modification of O'Neill and Roberts (1999), applied to discrete time models, and builds on the fact that animals remain on the same epidemic state for a long period. Briefly, for each chain successively one block of states $\mathbf{r}$ with maximum length M is chosen, and then one of three possible changes is proposed, as illustrated in Figure 6.3(a) on page 127. The proposed changes are: Add, Remove and Move. More precisely, in the "Add" step, we choose a period during which a given individual did not change their infection status and then proposed that for a subset of this period their infection status was reversed. Likewise in the "Remove" step we select an entire episode during which their infection status was unchanged and reverse this whole period, therefore joining the two neighbouring periods together. And a "Move" step move an endpoint of such a block. These changes propose a new vector $\mathbf{r}^*$, and the change is accepted with a probability that ensures the MCMC algorithm has the correct stationary distribution.

The efficiency of the algorithm depends on the size of the blocks that are proposed to be updated; if the blocks are too large then they no longer relate to the data and so are almost always rejected. The main advantage of this method is that the computational requirement is very small since the most of the hidden states are not updated. However, this can result in very slow mixing and therefore it needs to be run for more iterations in order to obtain independent samples.

6.4.3 Standard FFBS Gibbs update

For small number of chains, the whole hidden state process can be updated from its full conditional, $\pi\left(\mathbf{X}_{1:T}^{[1:C]} \mid \mathbf{Y}_{1:T}^{[1:C]}, \boldsymbol{\theta}\right)$, in a single block. This exact algorithm consists of translating the CHMM into an equivalent HMM with N^C states where $\mathbf{X}_t^{[1:C]} = \left(X_t^{[1]}, X_t^{[2]}, \ldots, X_t^{[C]}\right) \in \mathcal{X}_s^C = \{s_1, s_2, \ldots, s_N\}^C$ denote the state of the model at time t, as shown in Figure 6.3(c) on page 127.

Single block sampling can be achieved using the forward filtering backward sampling algorithm. This algorithm is based upon a forward recursion which calculates the filtered probabilities $\mathbb{P}\left(\mathbf{X}_t^{[1:C]} \mid \mathbf{Y}_{1:t}^{[1:C]}, \boldsymbol{\theta}\right)$. This is followed by a backward simulation step that first generates $\mathbf{X}_T^{[1:C]}$ from $\mathbb{P}\left(\mathbf{X}_T^{[1:C]} \mid \mathbf{Y}_{1:T}^{[1:C]}, \boldsymbol{\theta}\right)$ and then simulates the rest $\mathbf{X}_t^{[1:C]}$'s by progressing backwards, simulating in turn $\mathbf{X}_t^{[1:C]}$ from $\mathbb{P}\left(\mathbf{X}_t^{[1:C]} \mid \mathbf{X}_{t+1}^{[1:C]}, \mathbf{Y}_{1:t}^{[1:C]}, \boldsymbol{\theta}\right)$, for $t = T-1, T-2, \ldots, 1$. Hence, the full conditional distribution of the hidden states $\mathbf{X}_{1:T}^{[1:C]}$ given the observed data and the parameters of the Markov model is available in closed form. We denote this conditional distribution by $\pi_H\left(\mathbf{X}_{1:T}^{[1:C]} \mid \mathbf{Y}_{1:T}^{[1:C]}, \boldsymbol{\theta}\right)$. For more details about the algorithm see Chapter 3. Later in the chapter, we refer to this method as the fullFFBS.

The computational complexity of the forward-backward algorithm is $\mathcal{O}(TN^{2C})$. Thus, particularly for a reasonably large number of chains or possible states, this method will be computationally inefficient. This motivated us to derive a set of recursions to perform exact inference by using modified conditional forward-backward variables that can be calculated in time $\mathcal{O}(TN^2)$ for each chain.

6.4.4 Individual FFBS Gibbs updates

We propose a novel extension of the FFBS algorithm, where the hidden states are sampled individually per chain conditionally on the hidden states of the remaining chains, as opposed to the standard FFBS algorithm where sampling is done for all chains jointly. Figure 6.3(d) (see page 127) illustrates our proposed method, termed as iFFBS (individual FFBS) when the hidden states of the first chain are updated.

Under the conditional independence assumptions of our model, the full conditional distribution of $\mathbf{X}_{1:T}^{[c]}$, for each $c = 1, 2, \ldots, C$, can be factorised as:

$$\begin{aligned}
&\mathbb{P}\left(\mathbf{X}_{1:T}^{[c]} \mid \mathbf{X}_{1:T}^{[-c]}, \mathbf{Y}_{1:T}^{[1:C]}, \boldsymbol{\theta}\right) \\
&\qquad = \mathbb{P}\left(X_T^{[c]} \mid \mathbf{X}_{1:T}^{[-c]}, \mathbf{Y}_{1:T}^{[1:C]}, \boldsymbol{\theta}\right) \prod_{t=1}^{T-1} \mathbb{P}\left(X_t^{[c]} \mid \mathbf{X}_{t+1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \mathbf{Y}_{1:T}^{[1:C]}, \boldsymbol{\theta}\right) \\
&\qquad = \mathbb{P}\left(X_T^{[c]} \mid \mathbf{X}_{1:T}^{[-c]}, \mathbf{Y}_{1:T}^{[1:C]}, \boldsymbol{\theta}\right) \prod_{t=1}^{T-1} \mathbb{P}\left(X_t^{[c]} \mid X_{t+1}^{[c]}, \mathbf{X}_{1:t+1}^{[-c]}, \mathbf{Y}_{1:t}^{[c]}, \boldsymbol{\theta}\right)
\end{aligned}$$

where

$$\begin{aligned}
&\mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid X_{t+1}^{[c]} = x_{t+1}^{[c]}, \mathbf{X}_{1:t+1}^{[-c]}, \mathbf{Y}_{1:t}^{[c]}, \boldsymbol{\theta}\right) \\
&\qquad \propto \mathbb{P}\left(X_{t+1}^{[c]} = x_{t+1}^{[c]} \mid X_t^{[c]} = x_t^{[c]}, \mathbf{X}_{1:t+1}^{[-c]}, \boldsymbol{\theta}\right) \mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{1:t+1}^{[-c]}, \mathbf{Y}_{1:t}^{[c]}, \boldsymbol{\theta}\right) \\
&\qquad = \mathbb{P}\left(X_{t+1}^{[c]} = x_{t+1}^{[c]} \mid X_t^{[c]} = x_t^{[c]}, \mathbf{X}_t^{[-c]}, \boldsymbol{\theta}\right) \mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{1:t+1}^{[-c]}, \mathbf{Y}_{1:t}^{[c]}, \boldsymbol{\theta}\right) \\
&\qquad = P_{x_t^{[c]}, x_{t+1}^{[c]}, t+1}^{[c]} \mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{1:t+1}^{[-c]}, \mathbf{Y}_{1:t}^{[c]}, \boldsymbol{\theta}\right)
\end{aligned} \tag{6.5}$$

since the states of all chains a time $t+1$, $X_{t+1}^{[1]}, X_{t+1}^{[2]}, \ldots, X_{t+1}^{[C]}$, are independent conditional on the states of all chains at time t.

The rest of the calculation is concerned with determining the second mass function in Equation (6.5), which can be determined recursively for all t starting with

$t = 1$. We refer to this term as the modified conditional filtered probability. Similar to the standard forward backward procedure the forward recursion is initialised at $t = 1$ with:

$$
\begin{aligned}
&\mathbb{P}\left(X_1^{[c]} = x_1^{[c]} \mid \mathbf{X}_{1:2}^{[-c]}, \mathbf{Y}_1^{[c]}, \boldsymbol{\theta}\right) \\
&\qquad \propto \mathbb{P}\left(X_1^{[c]} = x_1^{[c]} \mid \boldsymbol{\theta}\right) f_{x_1^{[c]}}\left(y_1^{[c]} \mid \boldsymbol{\theta}\right) \left[\prod_{\substack{c'=1 \\ c' \neq c}}^{C} \mathbb{P}\left(X_2^{[c']} = x_2^{[c']} \mid X_1^{[c']} = x_1^{[c']}, \mathbf{X}_1^{[-c']}, \boldsymbol{\theta}\right)\right] \\
&\qquad = \nu^{[c]}_{x_1^{[c]}} \; f_{x_1^{[c]}}\left(y_1^{[c]} \mid \boldsymbol{\theta}\right) \quad \underbrace{\left[\prod_{\substack{c'=1 \\ c' \neq c}}^{C} P^{[c']}_{x_1^{[c']}, x_2^{[c']}, 2}\right]}_{\substack{\text{Transition probabilities of the} \\ \text{remaining chains at time } t=2}}
\end{aligned}
\tag{6.6}
$$

where the normalizing constant is the sum of the terms in Equation (6.6) as $X_1^{[c]}$ runs from s_1 to s_N. Then, for $t = 2, 3, \ldots, T-1$, we repeat the following steps:

Step 1. Compute the one-step ahead modified conditional predictive probabilities:

$$
\begin{aligned}
&\mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{1:t}^{[-c]}, \mathbf{Y}_{1:t-1}^{[c]}, \boldsymbol{\theta}\right) \\
&\qquad = \sum_{k \in \mathcal{X}_S} \mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid X_{t-1}^{[c]} = k, \mathbf{X}_{t-1}^{[-c]}, \boldsymbol{\theta}\right) \mathbb{P}\left(X_{t-1}^{[c]} = k \mid \mathbf{X}_{1:t}^{[-c]}, \mathbf{Y}_{1:t-1}^{[c]}, \boldsymbol{\theta}\right) \\
&\qquad = \sum_{k \in \mathcal{X}_s} P^{[c]}_{k, x_t^{[c]}, t} \; \mathbb{P}\left(X_{t-1}^{[c]} = k \mid \mathbf{X}_{1:t}^{[-c]}, \mathbf{Y}_{1:t-1}^{[c]}, \boldsymbol{\theta}\right)
\end{aligned}
\tag{6.7}
$$

Step 2. Compute the modified conditional filtered probabilities:

$$
\begin{aligned}
&\mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{1:t+1}^{[-c]} = \mathbf{x}_{1:t+1}^{[-c]}, \mathbf{Y}_{1:t}^{[c]}, \boldsymbol{\theta}\right) \\
&\qquad \propto \mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{1:t}^{[-c]}, \mathbf{Y}_{1:t-1}^{[c]}, \boldsymbol{\theta}\right) f_{x_t^{[c]}}\left(y_t^{[c]} \mid \boldsymbol{\theta}\right) \times \underbrace{\left[\prod_{\substack{c'=1 \\ c' \neq c}}^{C} P^{[c']}_{x_t^{[c']}, x_{t+1}^{[c']}, t+1}\right]}_{\substack{\text{Transition probabilities of the} \\ \text{remaining chains at time } t+1}}
\end{aligned}
\tag{6.8}
$$

where computing the normalising constant would require us to sum over the N possible values of $X_t^{[c]}$. Note that the last term in Equation (6.8) is calculated given $X_t^{[c]}$ and occurs due to $X_t^{[c]}$ connecting to $X_{t+1}^{[c']}$ in the graph of Figure 6.3(d), for $c' \neq c$.

The forward recursion is terminated at $t = T$ with:

$$\mathbb{P}\left(X_T^{[c]} = x_T^{[c]} \mid \mathbf{X}_{1:T}^{[-c]}, \mathbf{Y}_{1:T}^{[c]}, \boldsymbol{\theta}\right) = \frac{\mathbb{P}\left(X_T^{[c]} = x_T^{[c]} \mid \mathbf{X}_{1:T}^{[-c]}, \mathbf{Y}_{1:T-1}^{[c]}, \boldsymbol{\theta}\right) f_{x_T^{[c]}}\left(y_T^{[c]} \mid \boldsymbol{\theta}\right)}{\sum\limits_{k \in \mathcal{X}_s} \mathbb{P}\left(X_T^{[c]} = k \mid \mathbf{X}_{1:T}^{[-c]}, \mathbf{Y}_{1:T-1}^{[c]}, \boldsymbol{\theta}\right) f_k\left(y_T^{[c]} \mid \boldsymbol{\theta}\right)}.$$

Once the filtered probabilities have been calculated and stored in a forward sweep, the hidden states for a given chain c can be simulated in a backward sweep, starting with the simulation of a value for $X_T^{[c]}$ from the modified filtered probability at time T, $\mathbb{P}\left(X_T^{[c]} = x_T^{[c]} \mid \mathbf{X}_{1:T}^{[-c]}, \mathbf{Y}_{1:T}^{[c]}, \boldsymbol{\theta}\right)$. Then for $t = T-1, T-2, \ldots, 1$ we iteratively sample a value for $X_t^{[c]}$ given our simulated value for $X_{t+1}^{[c]}$, from:

$$\begin{aligned}\mathbb{P}&\left(X_t^{[c]} = x_t^{[c]} \mid X_{t+1}^{[c]} = x_{t+1}^{[c]}, \mathbf{X}_{1:t+1}^{[-c]}, \mathbf{Y}_{1:t}^{[c]}, \boldsymbol{\theta}\right) \\ &= \frac{P^{[c]}_{x_t^{[c]}, x_{t+1}^{[c]}, t+1}\, \mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{1:t+1}^{[-c]}, \mathbf{Y}_{1:t}^{[c]}, \boldsymbol{\theta}\right)}{\sum\limits_{k \,\in\, \mathcal{X}_s} P^{[c]}_{k, x_{t+1}^{[c]}, t+1}\, \mathbb{P}\left(X_t^{[c]} = k \mid \mathbf{X}_{1:t+1}^{[-c]}, \mathbf{Y}_{1:t}^{[c]}, \boldsymbol{\theta}\right)}.\end{aligned}$$

This forward-backward procedure provides the full conditional distribution of the hidden Markov chain c, denoted by $\pi_{H^{[c]}}\left(\mathbf{X}_{1:T}^{[c]} \mid \mathbf{Y}_{1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \boldsymbol{\theta}\right)$, in closed form. Therefore we can use a Gibbs sampler where each chain is updated conditional on the current values of the remaining chains, the model parameters and the observed data. The algorithm is presented in Algorithm 4.

Examining the computations, we see that the proposed iFFBS method requires TN^2 calculations for a single chain and hence a full sweep over all C chains requires TCN^2 as opposed to the TN^{2C} needed by fullFFBS. The key difference between the two methods, can be seen is Equation (6.7) where the sum for iFFBS is over N possible states whereas for fullFFBS the corresponding sum is over N^C terms. Another important difference is that evaluating the filtered probabilities of chain c at time $t < T$ for iFFBS involves the calculation of the transition probabilities of the remaining chains calculated at the next time point. Note that if these extra terms are omitted, then the iFFBS reduces to the standard FFBS. This latter approximation was used by Sherlock *et al.* (2013) for modelling interactions of different diseases. We call their method uncorrected-iFFBS. However, such an approximation requires an extra Metropolis Hastings step within the Gibbs sampling scheme to correct for the fact that the hidden states are not sampled from their full conditionals, as shown in the next section. Note that, failing to include the MH step may lead to poor behaviour of the resulting MCMC chains; an example is presented in Section E.1 of the Appendix.

```
Algorithm 4: MCMC algorithm for the Markov model with iFFBS method
Initialise: Draw $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and generate $\mathbf{X}_{1:T}^{[1:C]} \sim \pi(\mathbf{X}_{1:T}^{[1:C]} \mid \boldsymbol{\theta})$;
for $j$ = 1, 2, ..., $J$ do
    for $c$ = 1, 2, ..., $C$ do
        Draw $\mathbf{X}_{1:T}^{[c]} \sim \pi_{H^{[c]}}\left(\mathbf{X}_{1:T}^{[c]} \mid \mathbf{Y}_{1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \boldsymbol{\theta}\right)$ with iFFBS;
    end
    Perform suitable MCMC update to sample
     $\boldsymbol{\theta} \sim \pi\left(\boldsymbol{\theta} \mid \mathbf{Y}_{1:T}^{[1:C]}, \mathbf{X}_{1:T}^{[1:C]}\right)$;
end
```

6.4.5 Individual FFBS independence sampler

We assume that $\mathbf{X}_{t+1}^{[-c]}$ (the vector of the states of all other chains at time $t+1$ excluding the state of chain c, $X_{t+1}^{[c]}$) is dependent on $\mathbf{X}_t^{[-c]}$ but only weakly dependent on $X_t^{[c]}$. Motivated by this relation, each time we update chain c, we let $\mathbf{X}_{t+1}^{[-c]}$ to be independent of $X_t^{[c]}$, such that:

$$\mathbb{P}\left(X_{t+1}^{[c']} \mid \mathbf{X}_t^{[1:C]}, \boldsymbol{\theta}\right) = \mathbb{P}\left(X_{t+1}^{[c']} \mid \mathbf{X}_t^{[-c]}, \boldsymbol{\theta}\right), \quad \text{where } c' \neq c.$$

This assumption implies the Bayesian network shown in Figure 6.3(e), page 127. Given the assumption of independence, the product terms in Equations (6.6) and (6.8) cancel out, and so the modified conditional filtered probabilities reduce to:

$$\mathbb{P}\left(X_1^{[c]} = x_1^{[c]} \mid \mathbf{X}_{1:2}^{[-c]}, \mathbf{Y}_1^{[c]}, \boldsymbol{\theta}\right) = \frac{\nu_{x_1^{[c]}}^{[c]} f_{x_1^{[c]}}\left(y_1^{[c]} \mid \boldsymbol{\theta}\right)}{\sum\limits_{k \in \mathcal{X}_s} \nu_k^{[c]} f_k\left(y_1^{[c]} \mid \boldsymbol{\theta}\right)},$$

and

$$\mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{1:t+1}^{[-c]}, \mathbf{Y}_{1:t}^{[c]}, \boldsymbol{\theta}\right) = \frac{\mathbb{P}\left(X_t^{[c]} = x_t^{[c]} \mid \mathbf{X}_{1:t}^{[-c]}, \mathbf{Y}_{1:t-1}^{[c]}, \boldsymbol{\theta}\right) f_{x_t^{[c]}}\left(y_t^{[c]} \mid \boldsymbol{\theta}\right)}{\sum\limits_{k \in \mathcal{X}_s} \mathbb{P}\left(X_t^{[c]} = k \mid \mathbf{X}_{1:t}^{[-c]}, \mathbf{Y}_{1:t-1}^{[c]}, \boldsymbol{\theta}\right) f_k\left(y_t^{[c]} \mid \boldsymbol{\theta}\right)},$$

similar to the standard FFBS algorithm but with hidden state space $\mathcal{X}_s$ rather that $\mathcal{X}_s^C$. However, since we overlook some between chain dependencies our full conditionals are approximations of the true full conditionals. Therefore, we need to replace the Gibbs step with a Metropolis Hastings step to correct for the error of the approximation. We use an independence sampler with the approximated full conditionals as the proposal distributions. The detailed algorithm can be found in Algorithm 5. From here on, we refer to this proposed algorithm as MHiFFBS.

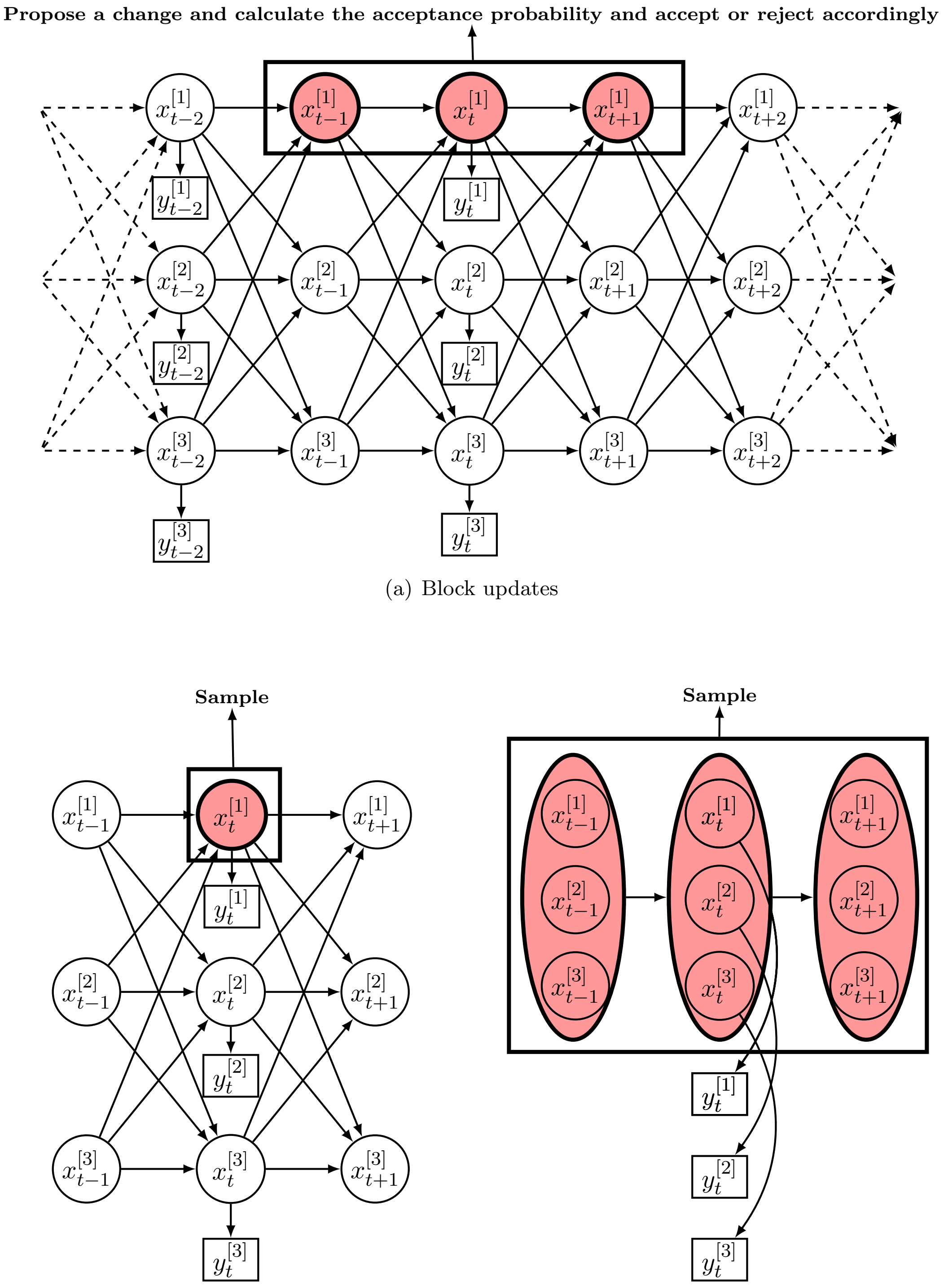
Algorithm 5: MCMC algorithm for the Markov model with MHiFFBS method

```
Initialize: Draw $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and generate $\mathbf{X}_{1:T}^{[1:C]} \sim \pi\left(\mathbf{X}_{1:T}^{[1:C]} \mid \boldsymbol{\theta}\right)$;
for $j = 1, 2, \dots, J$ do
    for $c = 1, 2, \dots, C$ do
        Propose $\mathbf{X}_{1:T}^{[c]*} \sim q\left(\mathbf{X}_{1:T}^{[c]} \mid \mathbf{Y}_{1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \boldsymbol{\theta}\right)$;
        Compute
          $a = \min\left(1, \frac{q\left(\mathbf{X}_{1:T}^{[c]} \mid \mathbf{Y}_{1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \boldsymbol{\theta}\right)}{q\left(\mathbf{X}_{1:T}^{[c]*} \mid \mathbf{Y}_{1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \boldsymbol{\theta}\right)} \times \frac{\pi\left(\mathbf{X}_{1:T}^{[c]*}, \mathbf{X}_{1:T}^{[-c]}, \boldsymbol{\theta} \mid \mathbf{Y}_{1:T}^{[1:C]}\right)}{\pi\left(\mathbf{X}_{1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \boldsymbol{\theta} \mid \mathbf{Y}_{1:T}^{[1:C]}\right)}\right)$;
        Draw $u \sim$ Uniform(0,1);
        if $u \leq a$ then
            Set $\mathbf{X}_{1:T}^{[c]} = \mathbf{X}_{1:T}^{[c]*}$;
        else
            Set $\mathbf{X}_{1:T}^{[c]} = \mathbf{X}_{1:T}^{[c]}$;
        end
    end
    Perform suitable MCMC update to sample
     $\boldsymbol{\theta} \sim \pi\left(\boldsymbol{\theta} \mid \mathbf{Y}_{1:T}^{[1:C]}, \mathbf{X}_{1:T}^{[1:C]}\right)$;
end
```

6.5 CHMMs and CHSMMs for modeling infection dynamics

Coupled hidden Markov models provide a natural way to model the transmission dynamics of an infectious disease where the coupling between different chains accounts for the interaction between individuals. In this section, we demonstrate how CHMMs can be embedded within an individual-based SIS epidemic model for the spread of infection through a household. Under this framework, the unobserved colonisation process corresponds to the hidden states of the CHMM and simulation can be done with the algorithms already described in Section 6.4.

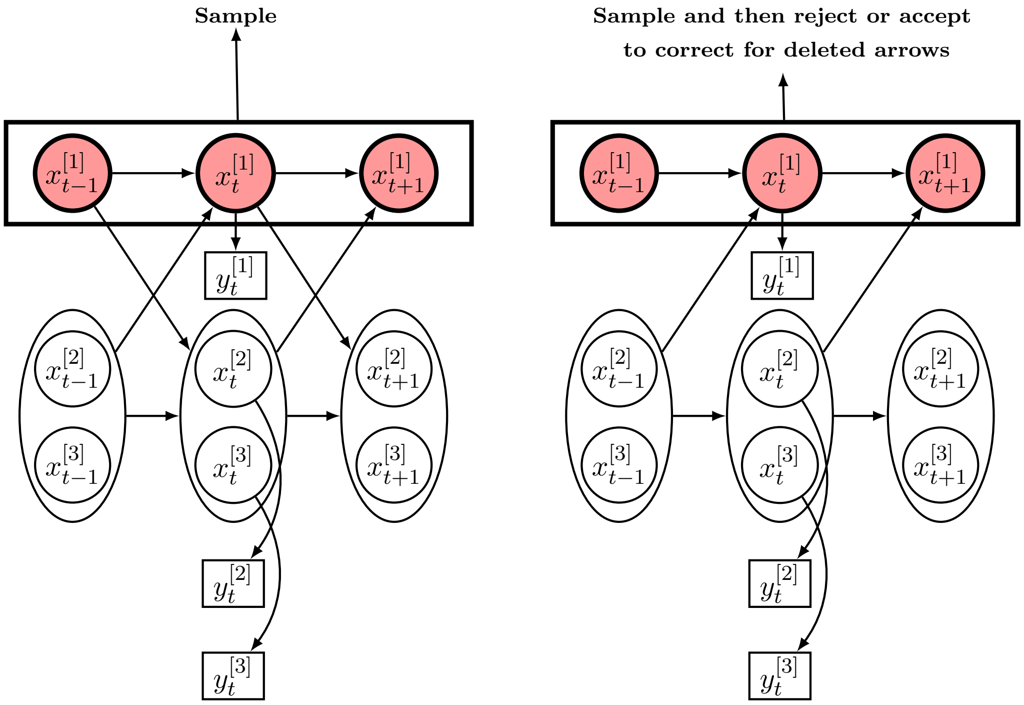
Figure 6.3: Strategies to model coupled hidden Markov models.



(a) Block updates

(b) Single site update

(c) fullFFBS



(d) iFFBS

(e) MHiFFBS

In Section 6.5.1 we discuss the SIS model with a Geometric distribution for the colonisation period, which yields a Markov model. In Section 6.5.2, we relax the Markovian assumption by allowing the duration to have a Negative Binomial distribution. This leads to a semi-Markov model in which the duration of colonisation depends on how long an individual has been infected.

6.5.1 Markov model

The SIS model with Markovian disease duration has been already described in Section 3.2 to which we refer the reader. The parameters of the model are sampled using either Gibbs or Hamiltonian Monte Carlo updates. The hidden infection states can be updated with any of the algorithms described in Section 6.4. Note that all methods are outlined for the case of a single pen. However, since we assume no interaction between pens it is straightforward to apply these methods in a problem with several pens. This independence assumption can be relaxed as shown in Section 6.7.2.

6.5.2 Semi-Markov model

In a departure from the previous Markov model, we assume that the time an individual remains infected has a two-parameter Negative Binomial distribution as described in Section 5.2.1. Bayesian inference for the semi-Markov model can proceed as follows. Regarding the update of the hidden states, the block proposals method can be applied without any modification. The single-site method can be done through the general Equation (6.4). For the fullFFBS and iFFBS methods the necessary Markov property is not valid, and therefore the two algorithms can not be applied directly. Therefore, we extend the methodology used before for updating the hidden states by considering an independence sampler within the MCMC algorithm. Our approach takes advantage of the availability of the full conditionals in the coupled hidden Markov model, by using them as a proposal in the update. More specifically, the proposal assumes a special case of the Negative Binomial distribution with $\kappa = 1$, equivalent to the Geometric, and therefore the efficiency of the algorithm depends on how close the real value of κ is to 1.

The full details of the extended algorithms for fullFFBS (called SM-fullFFBS) and iFFBS (called SM-iFFBS) are shown in Algorithm 6 and 7, respectively. In the same spirit, the MHiFFBS method can be also directly applied without modifications. As an alternative, one could use the method of Natarajan and Nevatia (2007) who proposed running the forward-backward recursions conditional on the durations vector $\boldsymbol{\zeta} = (\boldsymbol{\zeta}^{[1]}, \boldsymbol{\zeta}^{[2]}, \ldots, \boldsymbol{\zeta}^{[C]})$ and then marginalising over $\boldsymbol{\zeta}$. However, the computational cost of the recursions for coupled hidden semi-Markov models is much bigger than for CHMMs and increases as the observation period T grows. Therefore, we choose not to examine this option.

Similarly to the Markov model, we update the model parameters using either Gibbs or HMC steps. We now move to simulation studies.

Algorithm 6: MCMC algorithm for the semi-Markov model with SM-fullFFBS method

```
Initialise: Draw $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and generate $\mathbf{X}_{1:T}^{[1:C]} \sim \pi\left(\mathbf{X}_{1:T}^{[1:C]} \mid \boldsymbol{\theta}\right)$;
for $j$ = 1, 2, ..., $J$ do
    Propose $\mathbf{X}_{1:T}^{[1:C]\,*} \sim \pi_H\left(\mathbf{X}_{1:T}^{[1:C]} \mid \mathbf{Y}_{1:T}^{[1:C]}, \kappa = 1, \boldsymbol{\theta}_{-\kappa}\right)$ with FFBS;
    Compute $a = \min\left(1, \frac{\pi_H\left(\mathbf{X}_{1:T}^{[1:C]} | \mathbf{Y}_{1:T}^{[1:C]}, \kappa=1, \boldsymbol{\theta}^{-\kappa}\right)}{\pi_H\left(\mathbf{X}_{1:T}^{[1:C]\,*} | \mathbf{Y}_{1:T}^{[1:C]}, \kappa=1, \boldsymbol{\theta}^{-\kappa}\right)} \times \frac{\pi\left(\mathbf{X}_{1:T}^{[1:C]\,*}, \boldsymbol{\theta} | \mathbf{Y}_{1:T}^{[1:C]}\right)}{\pi\left(\mathbf{X}_{1:T}^{[1:C]}, \boldsymbol{\theta} | \mathbf{Y}_{1:T}^{[1:C]}\right)}\right)$;
    Draw $u \sim$ Uniform(0,1);
    if $u \leq a$ then
        Set $\mathbf{X}_{1:T}^{[1:C]} = \mathbf{X}_{1:T}^{[1:C]\,*}$;
    else
        Set $\mathbf{X}_{1:T}^{[1:C]} = \mathbf{X}_{1:T}^{[1:C]}$;
    end
    Perform suitable MCMC update to sample
     $\boldsymbol{\theta} \sim \pi\left(\boldsymbol{\theta} \mid \mathbf{Y}_{1:T}^{[1:C]}, \mathbf{X}_{1:T}^{[1:C]}\right)$;
end
```

Algorithm 7: MCMC algorithm for the semi-Markov model with SM-iFFBS method

```
Initialise: Draw $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and generate $\mathbf{X}_{1:T}^{[1:C]} \sim \pi\left(\mathbf{X}_{1:T}^{[1:C]} \mid \boldsymbol{\theta}\right)$;
for $j$ = 1, 2, ..., $J$ do
    for $c$ = 1, 2, ..., $C$ do
        Propose $\mathbf{X}_{1:T}^{[c]\,*} \sim \pi_{H^{[c]}}\left(\mathbf{X}_{1:T}^{[c]} \mid \mathbf{Y}_{1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \kappa = 1, \boldsymbol{\theta}_{-\kappa}\right)$ with
         iFFBS;
        Compute $a =$
         $\min\left(1, \frac{\pi_{H^{[c]}}\left(\mathbf{X}_{1:T}^{[c]} | \mathbf{Y}_{1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \kappa=1, \boldsymbol{\theta}^{-\kappa}\right)}{\pi_{H^{[c]}}\left(\mathbf{X}_{1:T}^{[c]\,*} | \mathbf{Y}_{1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \kappa=1, \boldsymbol{\theta}^{-\kappa}\right)} \times \frac{\pi\left(\mathbf{X}_{1:T}^{[c]\,*}, \mathbf{X}_{1:T}^{[-c]}, \boldsymbol{\theta} | \mathbf{Y}_{1:T}^{[1:C]}\right)}{\pi\left(\mathbf{X}_{1:T}^{[c]}, \mathbf{X}_{1:T}^{[-c]}, \boldsymbol{\theta} | \mathbf{Y}_{1:T}^{[1:C]}\right)}\right)$;
        Draw $u \sim$ Uniform(0,1);
        if $u \leq a$ then
            Set $\mathbf{X}_{1:T}^{[c]} = \mathbf{X}_{1:T}^{[c]\,*}$;
        else
            Set $\mathbf{X}_{1:T}^{[c]} = \mathbf{X}_{1:T}^{[c]}$;
        end
    end
    Perform suitable MCMC update to sample
     $\boldsymbol{\theta} \sim \pi\left(\boldsymbol{\theta} \mid \mathbf{Y}_{1:T}^{[1:C]}, \mathbf{X}_{1:T}^{[1:C]}\right)$;
end
```

6.6 Simulation studies

We perform a series of simulations to assess the efficiency of existing and proposed methods for updating the hidden infection states. Particular focus is given on how these methods are affected by dimensionality that is, when the total number of animals in the population and the study period increase. In Section 6.6.1 we study the Markov case whereas in Section 6.6.2 we apply the methods to data simulated from the semi-Markov model. Note that both throughout this Section as well as the real data analysis of Section 6.7 we allow for individual drop-outs during the study; all described methods can be easily modified for this scenario.

6.6.1 Markov model

The initial simulated dataset consists of observations from $P = 20$ pens, each containing $C = 8$ cattle as in the real dataset 1. Moreover, the study period is set to $T = 99$ days. First, we generate the hidden colonisation states according to the model defined in Equation (3.1) of Chapter 3, using the same parameter values as estimated in Section 3.6.1. In particular, the simulated data are generated with an external colonisation parameter α equal to 0.009, within-pen colonisation parameter β equal to 0.01, mean colonisation period m equal to 9 days and initial probability of colonisation set to $\nu = 0.1$. We then generate RAMS and faecal samples from the population at intervals according to the actual sampling frame employed in dataset 1. Finally, the RAMS and faecal test sensitivities are assumed to be 0.8 and 0.5, respectively.

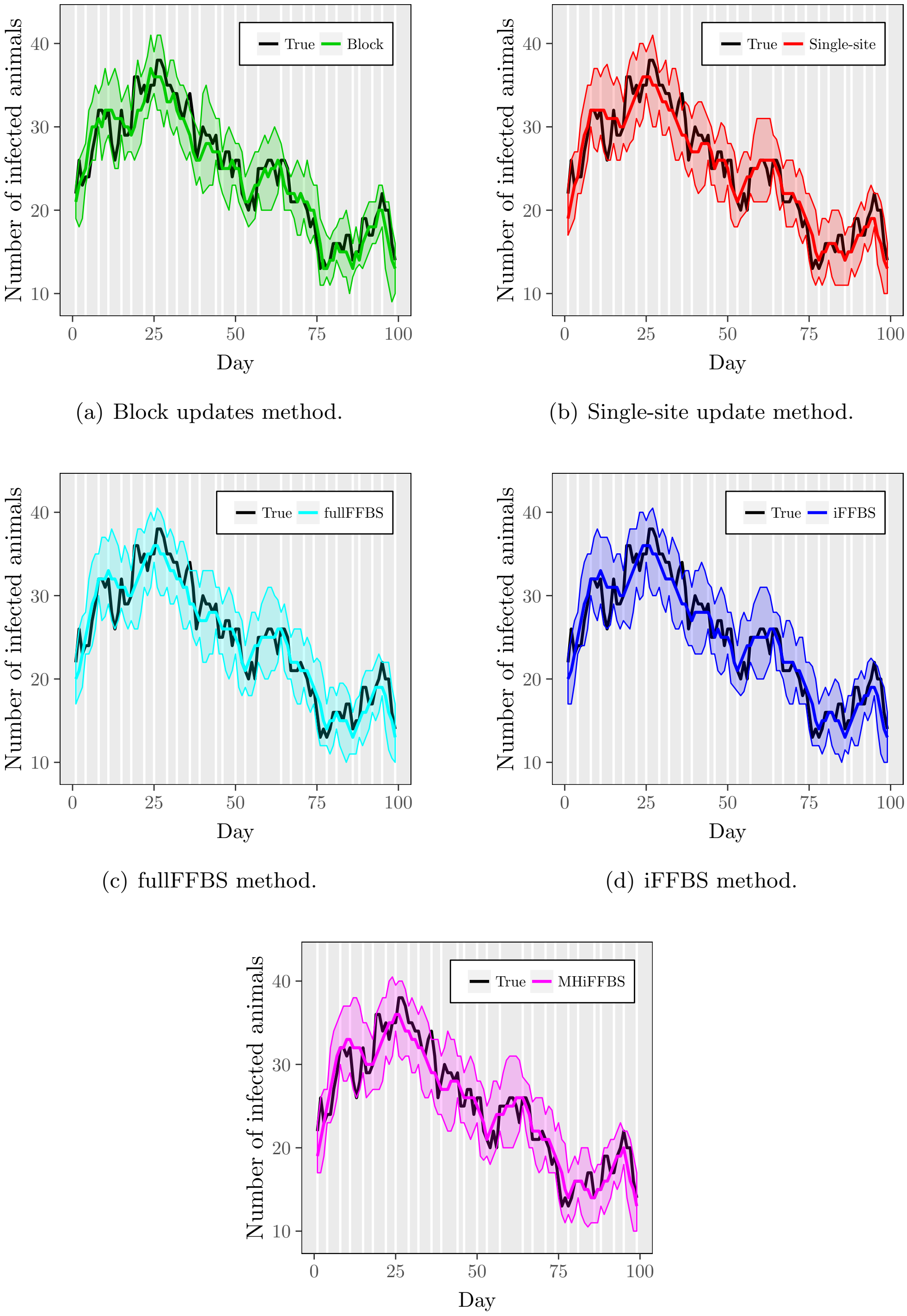
For the unknown parameters in the Markov model, the prior distributions are specified as follows: $\alpha, \beta \sim$ Ga(1,1), $m - 1 \sim$ Ga(0.01, 0.01) and $\nu, \theta_R, \theta_F \sim$ Beta(1,1). We draw samples from the joint posterior of the hidden states and model parameters with the MCMC scheme described earlier in the chapter, using all possible methods for updating the hidden states. The model parameters ν, θ_R and θ_F are updated using Gibbs steps and the remaining parameters are updated jointly using the Hamiltonian Monte Carlo, as in Section 3.4. For each method, we run the algorithm for 11,000 iterations, removing the first 1,000 as a burn-in. Each procedure is repeated 20 times to provide an empirical Monte Carlo estimate of the variation in each approach.

Figure 6.4 shows the estimated number of infected individuals over time, along with 95% credible intervals, as obtained from one of the 20 runs. We generally see that the 5 methods provide almost identical results and all of them contain the true total number of infected individuals within the credible intervals. Therefore, a comparison of the different approaches can be based on the mixing properties and the required computational effort of each. Mixing can measured in terms of autocorrelation of the Markov chains whereas the computational effort is given by the total time required for one iteration of the MCMC. In the following results we choose our summary statistic to be the total number of infected individuals $I = \sum_{p=1}^{P}\sum_{c=1}^{C}\sum_{t=1}^{T} x_t^{[c,p]}$, in order capture the information over all T periods of the study.

In the left panel of Figure 6.5 we see the autocorrelation function (ACF) function for I, averaged across the 20 different runs in each method. We see that the fullFFBS, iFFBS and MHiFFBS methods have very good mixing properties since the autocorrelation function drops rapidly as a function of time. On the contrary, both block proposals and single-site methods produce highly correlated samples with the ACF function being greater than zero even after 30 iterations of the MCMC. Both findings are expected as discussed in Section 6.4. For the block proposals method slow mixing is due to only few states being updated at each iteration of the MCMC; for the single-site method slow mixing is caused by the strong correlation between hidden states. As far as the running time is concerned, we find that block proposals method is the fastest as can be seen from the right panel of Figure 6.5. The computationally most demanding method is fullFFBS due to the summation over all possible 2^8 states that needs to be calculated.

Nevertheless, direct comparison of either ACF functions or CPU time is not indicative of computational efficiency on its own. For example, we see that block proposals are optimal in terms of CPU time per iteration, but results in much slower mixing as compared to the other methods. Therefore, a measure of computational efficiency needs to account for both features discussed before. We use the relative speed which is defined as following. First, for each method we calculate the time normalised effective sample size (tESS), the ratio of effective sample size (ESS) from 10,000 MCMC iterations and the CPU time required per iteration. Then, we divide the tESS of each method with the worst observed tESS to obtain the relative speed. Hence, the relative speed has a minimum value of 1 which corresponds to the computationally least efficient method, and any number bigger than 1 reflects the gains using a particular method compared to the worst. In the left panel of Figure 6.6 we show the relative speed of each method as obtained from the 20 different runs. We observe that among competing methods, the proposed iFFBS methods is the one that best combines the desired properties of mixing and computational speed, followed by the fullFFBS and the proposed MHiFFBS methods. Block proposals method is the least efficient method as it has a relative speed of 1 in all 20 replicates.

Figure 6.4: Median posterior number of infected individuals for the Markov model. Black solid lines represent the true values. Shaded regions represent the 95% credible intervals. White vertical lines represent the days where samples were collected.



(a) Block updates method.

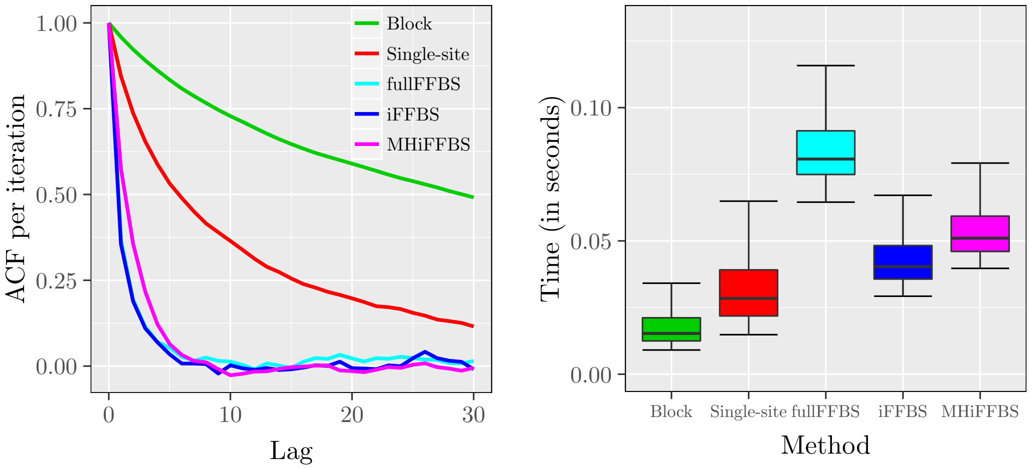
(b) Single-site update method.

(c) fullFFBS method.

(d) iFFBS method.

(e) MHiFFBS method.

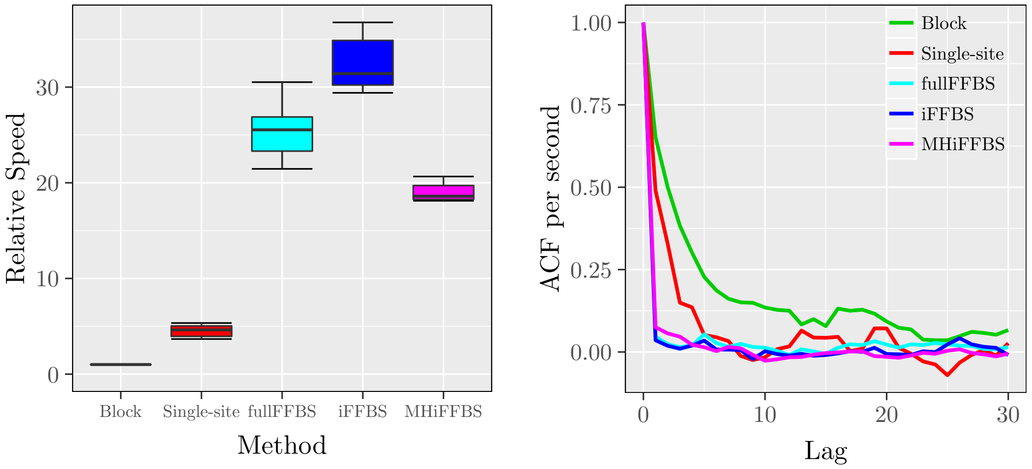
Figure 6.5: Autocorrelation function of I (left) and CPU time per iteration (right) for the Markov epidemic model. ACF plots in the left panel are the average across 20 replicates. Quantiles in the right panel are obtained from the same 20 runs.



(a) ACF per iteration.

(b) CPU time per iteration.

Figure 6.6: Relative speed (left) and ACF per second for I (right) for the Markov epidemic model. Quantiles in the left panel are obtained from 20 runs. ACF plots in the right panel are the average of the same 20 runs.



(a) Relative speed.

(b) ACF per second.

This finding is confirmed in the right panel of Figure 6.6 where we show the ACF per second.

In the next set of simulations we study how computation time is affected as we vary the total population size. To do so we use our initial simulation setting and generate one dataset for different values of C, $C = 3, 4, \ldots, 11$. Figure 6.7 illustrates the time taken per iteration of the five different methods as it varies with the number of animals in a pen. We see that the method that is affected the most is the fullFFBS, for which computational time grows exponentially with C. The other methods are only affected linearly when C increases. As before, we assess computational efficiency with the relative speed. Results are summarised in Table 6.1(a). Note that despite it being the computationally most efficient for small $C = 3, 4, \ldots, 7$, the performance of FFBS drops with C and eventually for $C = 11$ is found to have the lowest relative speed. For $C > 7$, the iFFBS method outperforms the remaining methods. In order to study the influence of the study length had on the performance of each method, we repeat our study for different values of T. Results are given in Table 6.1(b). Again, the iFFBS method is the one that scores higher in terms of relative speed, followed in order by fullFFBS, MHiFFBS, single-site and block proposals methods.

FIGURE 6.7: CPU time per iteration as a function of the total number of cattle per pen C, for the Markov epidemic model. Inner panel is a zoom-in.

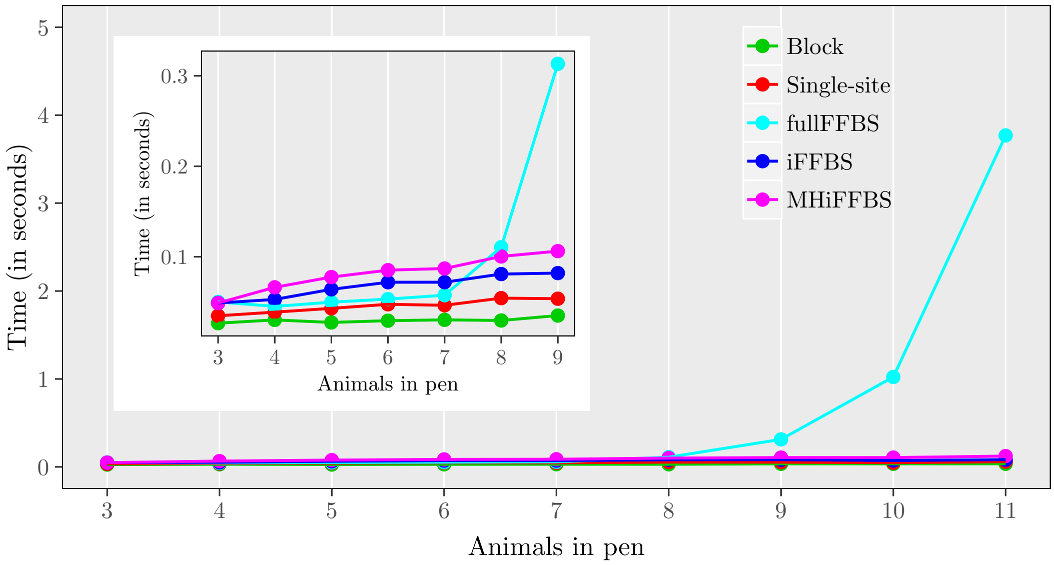


Table 6.1: Relative speed comparison of the five methods in the Markov model (a) as a function of the total number of cattle per pen C and (b) as a function of study period T. Bold entries represent the most efficient method in each setup.

(a) Varying number of animals over a 14-week period study.

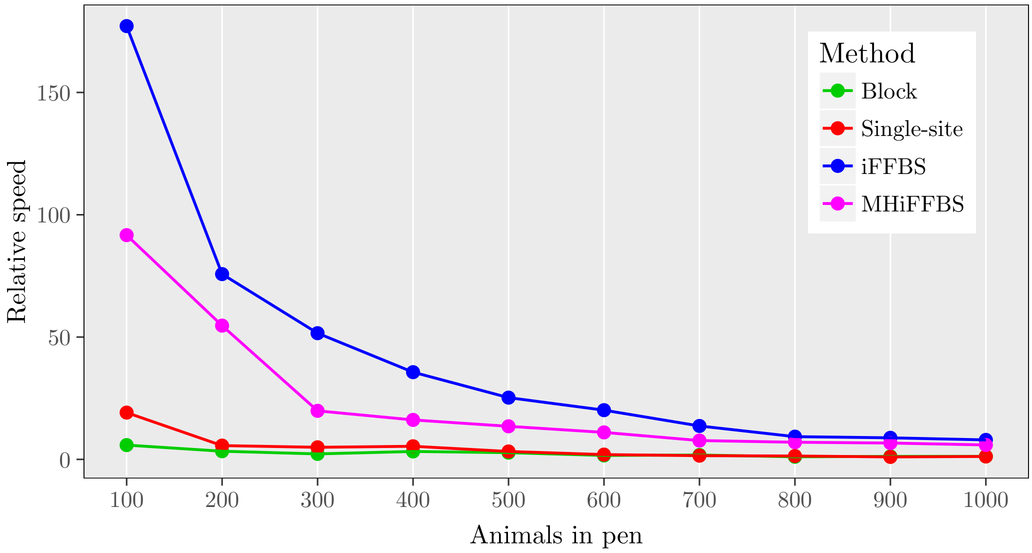
Number of	**Methods**				
animals	**Block**	**Single-site**	**fullFFBS**	**iFFBS**	**MHiFFBS**
3	5.10	11.79	**103.43**	96.08	46.19
4	3.87	11.17	**86.72**	85.12	43.42
5	4.33	9.68	**84.23**	73.60	39.33
6	3.47	7.64	**82.55**	61.51	39.38
7	3.19	9.19	**78.35**	63.18	37.01
8	3.37	7.30	41.33	**57.92**	29.46
9	2.85	6.96	13.92	**51.42**	26.28
10	2.61	7.59	4.38	**55.68**	28.34
11	2.15	5.96	1.00	**49.37**	22.05

(b) Varying study period with 8 number of animals.

Study	**Methods**				
period	**Block**	**Single-site**	**fullFFBS**	**iFFBS**	**MHiFFBS**
4 weeks	21.03	31.28	122.42	**136.00**	81.14
9 weeks	10.44	20.10	74.26	**126.99**	64.73
14 weeks	5.86	13.36	56.10	**95.25**	53.81
19 weeks	2.84	9.60	44.67	**77.78**	41.61
24 weeks	2.31	6.82	41.27	**52.48**	31.02
29 weeks	2.00	6.14	34.62	**50.37**	24.85
34 weeks	1.77	5.62	29.85	**43.66**	21.30
39 weeks	1.00	4.85	25.98	**38.91**	18.10
44 weeks	1.18	5.02	21.65	**39.83**	14.22

In our simulations so far we have evaluated the performance of the five methods for data of moderate dimensionality; however, an interesting question one could ask is how well the algorithms scale for large datasets. We investigate this question by analysing synthetic datasets with large C. In this setup, application of the fullFFBS method is computationally prohibitive and hence is not included in the simulations. We visualise the results in Figure 6.8. As before, the iFFBS outperforms the other methods whereas the least efficient is the block updates with a relative speed equal to 1 in all scenarios. The gains of using the iFFBS algorithm are higher in the first scenario with 100 animals per pen, where the method has a relative speed of 177.13. However, the differences in the computational efficiency among methods are less profound as the total number of individuals per pen increases. For example, in the last scenario ($C = 1000$) the iFFBS algorithm has relative speed 7.97.

FIGURE 6.8: Relative speed with large datasets for the Markov epidemic model.



6.6.2 Semi-Markov model

In this section we repeat the analyses of Section 6.6.1, this time for the semi-Markov model. The extra parameter κ is set to 1.6 as estimated in the analysis of dataset 1 by Spencer *et al.* (2015). We give an uninformative Ga(0.01,0.01) prior to κ and estimate it with the remaining parameters in the MCMC as described in Section 5.2.2. We now summarise the findings.

We find little difference in the estimated number of infected individuals across methods and once again these estimates are close to the real values (Figure 6.9). Their difference is highlighted in Figure 6.10(a) which compares CPU timings and relative speeds. In this semi-Markov model, both block updates and MHiFFBS methods can be applied without any modification and therefore require the same time per iteration; the remaining methods are slowed due to the modifications explained in Section 6.5.2 (also see Figure 6.10, left panel). In terms of relative speed, MHiFFBS has a slightly higher median compared to SM-iFFBS which is second best, followed by SM-fullFFBS, block proposals and single-site methods (Figure 6.10, right panel); however the best two have overlapping credible intervals. Comparing Figure 6.10(b) with Figure 6.6(a) we conclude that the gains of using the proposed algorithms drop when we move from the Markov to the more complex semi-Markov model. For SM-iFFBS this fact is due to the extra MH step introduced within the sampler.

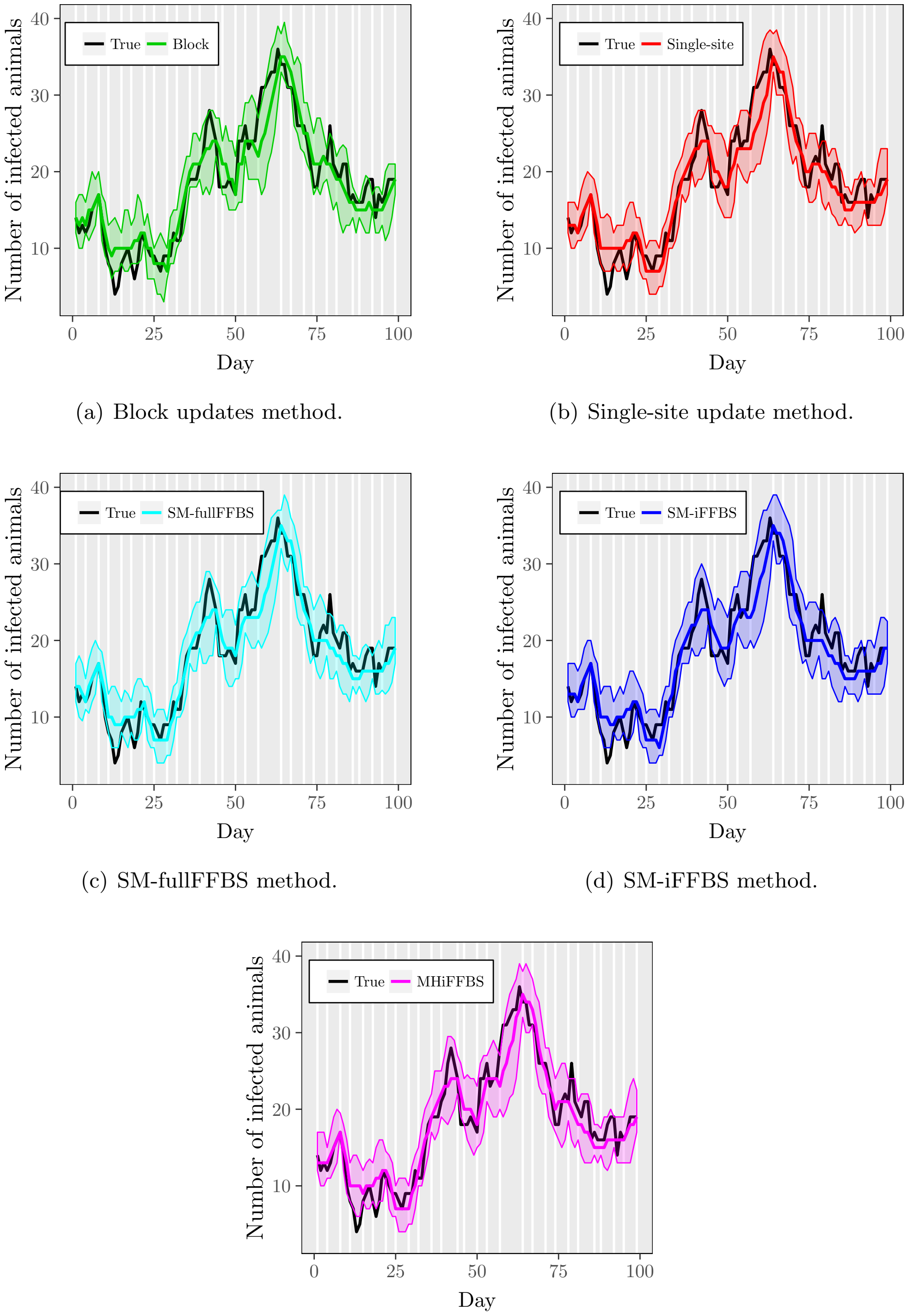
Results of relative speed for several values of C and T are shown in Table 6.2(a) and Table 6.2(b) respectively. For this model the SM-iFFBS approach has similar performance to the MHiFFBS. MHiFFBS has the highest relative speed in 14 out of the 18 simulated datasets whereas SM-iFFBS is the most efficient in 3 out of 18 occasions; nevertheless the differences are small in most of the occasions. Another interesting observation is that block updates method now produces better relative speed than the single-site method in 16 out of 18 simulations. Finally, for large datasets once again we observe superiority of the two proposed methods in relative speed (Figure 6.11) but gains are less substantial than the Markov case (Figure 6.8).

Table 6.2: Relative speed comparison of the five methods in the semi-Markov model (a) as a function of the total number of cattle per pen C and (b) as a function of study period T. Bold entries represent the most efficient method in each setup.

(a) Varying number of animals over a 14-week period study.

Number of	**Methods**				
animals	**Block**	**Single-site**	**SM-fullFFBS**	**SM-iFFBS**	**MHiFFBS**
3	47.75	52.19	**242.10**	201.93	240.22
4	47.74	39.97	197.83	199.29	**237.70**
5	38.88	35.96	171.65	193.20	**228.36**
6	33.42	29.00	94.49	191.65	**221.21**
7	28.96	22.91	83.05	**184.85**	174.82
8	27.70	21.43	43.53	141.77	**144.17**
9	22.61	18.12	13.61	119.34	**132.85**
10	21.88	13.75	4.59	102.32	**113.36**
11	18.82	14.04	1.00	100.94	**106.28**

Figure 6.9: Median posterior number of infected individuals for the semi-Markov model. Black solid lines represent the true values. Shaded regions represent the 95% credible intervals. White vertical lines represent the days where samples were collected.



(a) Block updates method.

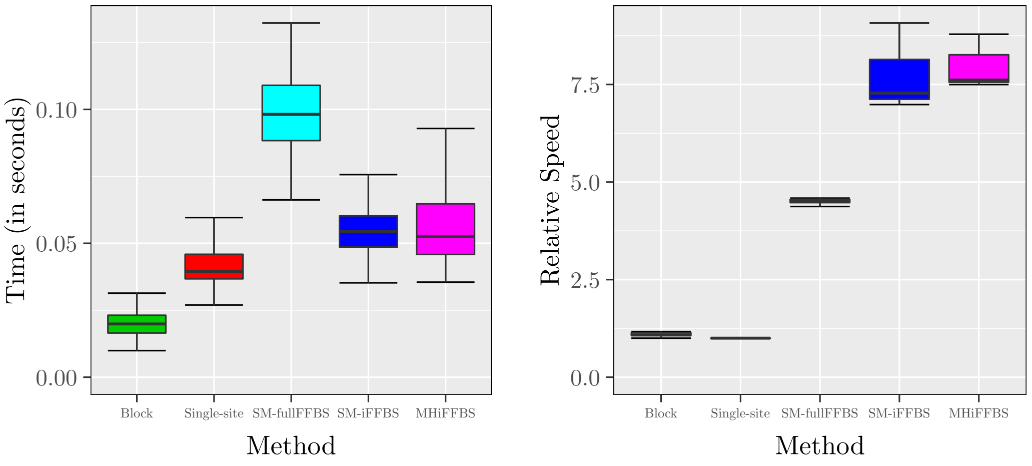
(b) Single-site update method.

(c) SM-fullFFBS method.

(d) SM-iFFBS method.

(e) MHiFFBS method.

Figure 6.10: CPU time per iteration (left) and relative speed (right) for the semi-Markov epidemic model. Quantiles in both left and right panels are obtained from 20 different replicates.



(a) Time per iteration.

(b) Relative speed.

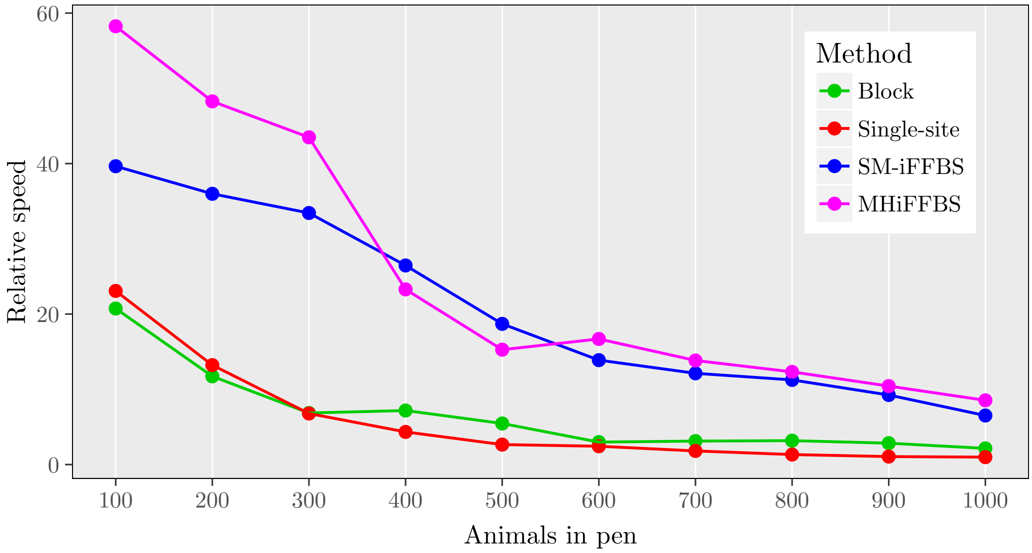
(b) Varying study period with 8 animals per pen.

Study	**Methods**				
period	**Block**	**Single-site**	**SM-fullFFBS**	**SM-iFFBS**	**MHiFFBS**
4 weeks	16.50	14.94	18.06	39.00	**55.18**
9 weeks	8.98	6.64	16.36	28.73	**44.65**
14 weeks	5.72	4.16	12.91	**26.30**	26.13
19 weeks	4.89	2.78	6.47	22.13	**23.54**
24 weeks	2.68	2.05	4.68	17.89	**20.35**
29 weeks	2.51	1.72	4.51	14.24	**16.55**
34 weeks	1.96	1.47	3.04	**11.60**	11.47
39 weeks	1.78	1.28	2.64	10.00	**10.03**
44 weeks	1.13	1.00	2.50	9.08	**9.47**

6.7 Applications

In this section we use the methods described throughout the chapter for the analysis of the datasets presented in Section 2.2. The results that we present focus on the comparison between the different algorithms rather than drawing conclusions regarding the dynamics of *E. coli* O157:H7. In Section 6.7.1 we discuss the first

Figure 6.11: Relative speed for large datasets for semi-Markov models.



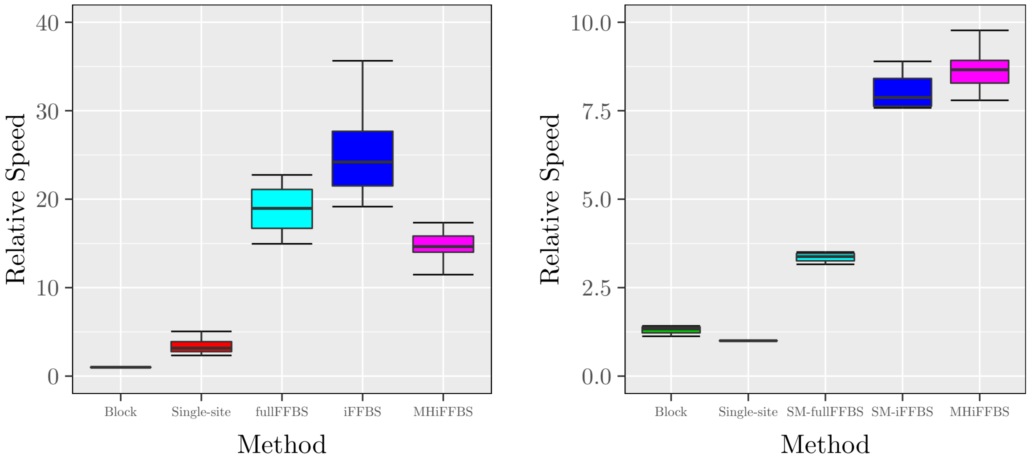
dataset and in Section 6.7.2 we analyse the second.

6.7.1 Dataset 1

We fit both the Markov and semi-Markov models. Priors specifications are identical to the ones used for the analysis of the simulated data in Sections 6.6.1 and 6.6.2. MCMC is run for 11,000 iterations with the first 1,000 discarded as burn-in and the remaining 10,000 used for the calculation of the relative efficiency. To get an estimate of the Monte Carlo variability of the efficiency measure we run 20 chains per method, with different starting values.

Posterior summaries of the model parameters are presented in Table 6.3 where we observe that the estimates obtained from the different methods are in close agreement. Note that we show the summaries only for the proposed iFFBS and MHiFFBS methods since the results obtained from the remaining methods are similar. The comparison of computational efficiency yields conclusions which are analogous to the ones reached when comparing performance on simulated data. For the Markov model we find a median relative speed of 1 for block proposals method, single-site method has 3.16, MHiFFBS has 14.65, while the fullFFBS and iFFBS are the most efficient with relative speeds of 18.96 and 24.27 respectively. For the semi-Markov the medians are 1 for single-site method, 1.33 for block updates, 3.37 for SM-fullFFBS, 7.87 for the SM-iFFBS and finally 8.65 for MHiFFBS. Results are

Figure 6.12: Relative speed comparison of methods when applied for the analysis of dataset 1. (a) Results for the Markov model. (b) Results for the semi-Markov model. Quantiles are obtained from 20 different replicates.



(a) Markov model.

(b) Semi-Markov model.

shown in Figure 6.12.

Table 6.3: Posterior summaries for the parameters of both Markov and semi-Markov epidemic models, fit to dataset 1. The two methods presented are iFFBS and MHiFFBS. S.d. indicates standard deviation.

Symbol	Geometric				Negative Binomial			
	iFFBS		MHiFFBS		SM-iFFBS		MHiFFBS	
	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.
α	0.009	0.001	0.009	0.001	0.008	0.001	0.008	0.001
β	0.011	0.002	0.011	0.002	0.010	0.002	0.010	0.002
m	9.365	0.743	9.227	0.789	10.033	0.837	10.061	0.822
κ	–	–	–	–	1.680	0.485	1.659	0.456
ν	0.098	0.025	0.099	0.025	0.099	0.026	0.098	0.025
θ_R	0.777	0.022	0.775	0.024	0.772	0.024	0.774	0.023
θ_F	0.466	0.022	0.464	0.022	0.462	0.023	0.464	0.022
I	2279	50.87	2286	58.44	2296	58.27	2287	57.74

6.7.2 Dataset 2

For this section, we relax the assumption of pen independence thus accounting for interaction between neighbours. In particular, we consider a Markov model that allows for *E. coli* O157:H7 transmission through shared waterers, since in Section 5.4.2.2 we found that this model is best supported by the data compared to other models allowing for between neighbour transmission.

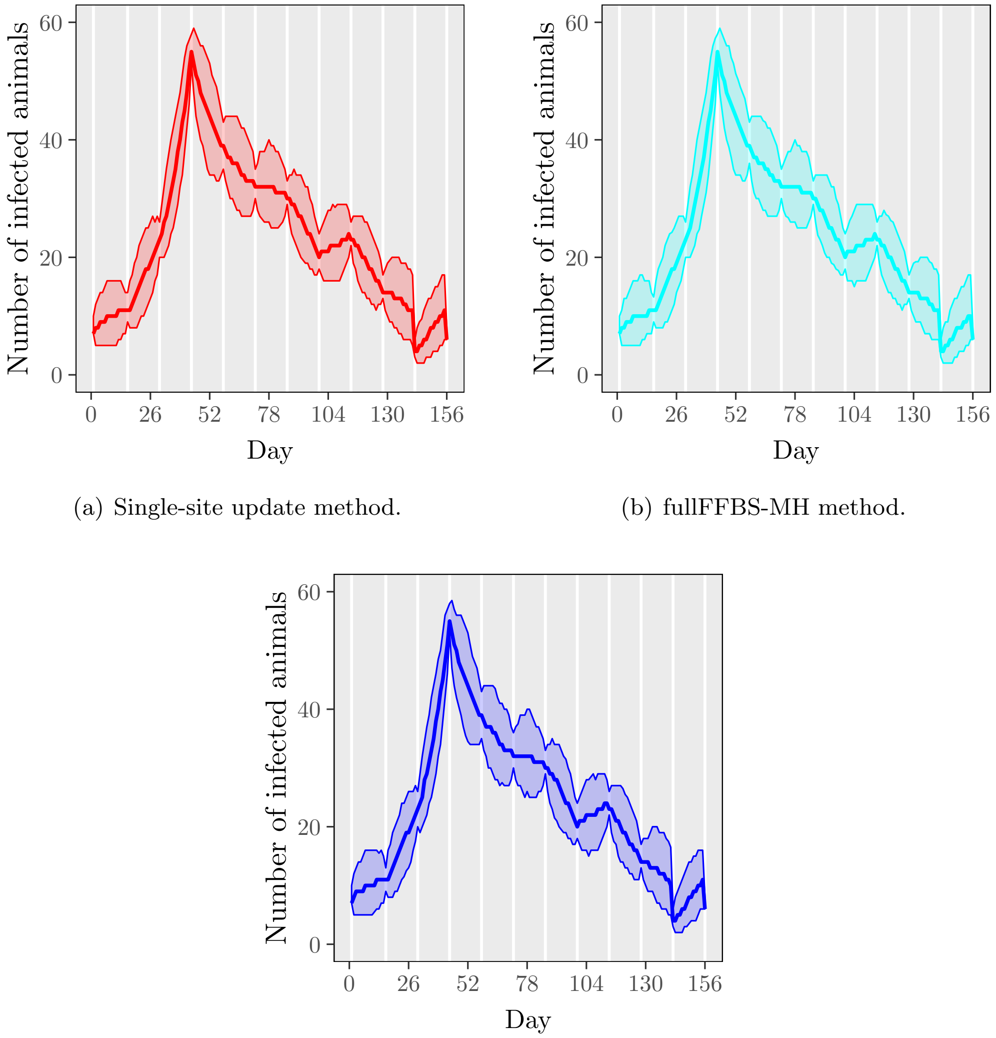
The methods that we consider are single-site updates and iFFBS; some additional terms appear in the full conditional distributions to account for interaction between animals in different pens. In iFFBS, updates for a hidden chain c are done conditionally not only on the chains of the remaining subjects in the pen but also conditionally on the chains of individuals in the neighbouring pens. As a result, the modified filtered probabilities additionally include the transition probabilities of subjects in neighbouring pens. Application of the fullFFBS is not possible within a reasonable amount of time since there is a summation over 2^{14} possible states that needs to be evaluated. Instead, we apply the fullFFBS-MH method which has been already used for this dataset in Section 5.4.2.2.

For the model parameters, we assign independent priors: $\alpha, \beta, \eta \sim$ Gamma $\big(1, 1\big)$, $m - 1 \sim \text{Ga}(0.01, 0.01)$ and $\nu \sim \text{Beta}(1, 1)$. As before, we run the MCMC for 11,000 iterations, using the last 10,000 to do the comparisons. Estimates obtained from the three methods are almost identical to the third decimal place (Table 6.4) and so is the estimated number of infected animals per day (Figure 6.13). The iFFBS method outperforms the single-site method in computational efficiency as indicated by the median relative speed of 15.61 and fullFFBS-MH as indicated by the median relative speed of 2.56.

Table 6.4: Posterior summaries for the parameters of the Markov epidemic model, fit to dataset 2. The three methods considered are Dong's, iFFBS and fullFFBS using a Metropolis Hastings step (fullFFBS-MH) as described Section 5.4.2.2. S.d. indicates standard deviation.

Parameter	**Method**					
	Single-site		**iFFBS**		**fullFFBS-MH**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
α	0.0032	0.0008	0.0034	0.0008	0.0035	0.0009
β	0.0099	0.0016	0.0097	0.0017	0.0099	0.0018
δ	0.0014	0.0009	0.0015	0.0009	0.0010	0.0007
ν	0.0495	0.0175	0.0494	0.0176	0.0493	0.0175

Figure 6.13: Median posterior number of infected studies per day in dataset 2, obtained fitting the Markov epidemic model. The three methods considered are iFFBS (left, red) and Dong's (right, blue). Shaded areas represent the 95% credible intervals. White vertical lines represent the days for which samples were taken.



(a) Single-site update method.

(b) fullFFBS-MH method.

(c) iFFBS method.

6.8 Discussion

In this chapter, we have considered the problem of Bayesian estimation of the hidden states in coupled hidden Markov models, an extension of the classical hidden Markov model, which allows for interactions between hidden states of each individual. In particular, we have reviewed existing methods in the field that include block proposals method, single-site method and the FFBS algorithm, and introduced two new approaches, the iFFBS and MHiFFBS algorithms. We have also extended the methods to the coupled hidden semi-Markov model in which the hidden process can remain in a given state for a non-memoryless duration. The utility of all methods, including existing and proposed, has been demonstrated in the context of modelling the dynamics of an infectious disease where we have assumed both a Markov and a semi-Markov model for the duration of the disease.

In our extended simulation studies we have demonstrated the merits of the proposed methods compared to the existing methods that have been considered. Our methods balance the desired properties of good mixing and low CPU time and thus prove to be computationally more efficient than previous methods in the field, providing at the same time estimates of the same quality. The findings are stronger for the Markov model but also hold in the semi-Markov case. Additionally, we have also demonstrated that the proposed methods scale well for big datasets, as opposed to the standard FFBS algorithm which cannot be applied when the number of chains in the CHMM is growing.

Finally, we have used the iFFBS and MHiFFBS for the analysis of two real datasets concerning the transmission dynamics of *E. coli* O157:H7 in cattle and again found that our approaches perform better in terms of computational efficiency. Furthermore, we have demonstrated how iFFBS can be used for inference in epidemic models allowing for interactions between neighbouring pens.

There are several ways in which the proposed methodologies can be extended. In the current approach, we update the states of a single chain given the rest. One idea is to apply a block update scheme, where a small subset of chains is jointly sampled from its full conditional. This would be particularly effective when there is some population structure such as households that could be used to define the blocks. Another possibility would be to only update k randomly selected chains per iteration instead of updating all chains. That may result into a substantial computational speedup and a particularly interesting question is how an optimal k might be chosen. Some initial results indicate that for small populations ($C < 200$), this approach leads to no substantial increase in relative speed; however, we expect the gains to be higher for larger population sizes. As an alternative, one could determine the probability of updating an individual's chain based on a pilot iFFBS run.

Chapter 7

Multi-State Markov Model For Longitudinal Data With Misclassification

7.1 Introduction

In our work so far, the true infection status of an individual is viewed as a binary process; one can either carry the disease or not. This is a common assumption in many epidemiological studies (Auranen *et al.*, 2000; Melegaro *et al.*, 2004; Matthews *et al.*, 2006b; Spencer *et al.*, 2015, for example). However, it is often the case that additional information exists regarding the serotype in which a disease appears. In such cases, it is reasonable to examine whether there is appreciable heterogeneity between the different serotypes in, for example, their transmissibility or the duration for which each serotype remains at the host. Additionally, we would like to address the question of between serotype competition, that is if carriage of certain type reduces the possibility of being colonised by a different type. Such knowledge can further our understanding of the epidemiology of an infectious disease and aid the policy decision making during an epidemic.

Nevertheless, parameter estimation in a multi-type disease context may be challenging due to identifiability issues, which occur due to several serotype-specific parameters that need to be estimated. Another problem is the large amount of missing data as a consequence of the sparse sampling intervals that are often used. A solution to the former can be given by grouping multiple serotypes into one class in order to simplify the model. Some examples include Cauchemez *et al.* (2006) who group serotypes as vaccine and non-vaccine, Erästö *et al.* (2012) who classify serotypes according to their frequency in the real data and Melegaro *et al.* (2007) who used a separate model for each serotype. More recently, Worby *et al.* (2016) used genome sequence information to classify the isolates into genetically similar groups. In all of these approaches, the problem of missing data is dealt either by extending the Bayesian data augmentation framework proposed by O'Neill and Roberts (1999) (Cauchemez *et al.*, 2006; Erästö *et al.*, 2012; Worby *et al.*, 2016), or by adopting a maximum profile likelihood approach (Melegaro *et al.*, 2007).

An additional complication arises from the fact that serotype information often relies on diagnostic tests which suffer from low sensitivity. As a result, several carriage incidents remain undetected or might be recorded as a wrong serotype. Most of the above methods assume that test results are observed without error and hence do not allow the possibility of false negative outcomes. The exception is Worby *et al.* (2016) who estimate a common test sensitivity for all groups. However, their model does not allow for between-serotype competition. Therefore, separating misclassification from changing serotype is an extremely challenging statistical problem that has not been solved successfully.

In this chapter we extend our previous modelling framework to allow for carriage of *E. coli* with multiple serotypes, available for a longitudinal study of the disease in Canada in 2003. The problem of the missing carriage states is tackled with Bayesian data augmentation where we also allow for the possibility of type misclassification. The chapter is structured as follows. We formulate our model in Section 7.2 and in Section 7.3 we describe the algorithm which is used for posterior inference. Performance of our method is assessed on simulated data under different scenarios in Section 7.4. In Section 7.5 we apply the proposed methodology to the dataset described in Section 2.4 and in Section 7.6 we conclude with a discussion.

7.2 Model

In our study, the diagnostic tests that where conducted to check the presence of *E. coli* O157:H7 in cattle are imperfect; the sensitivities of these techniques may be as low as 50% and thus some colonised individuals remain undetected. In addition, the PFGE method used for identifying serotypes in samples has also less than perfect sensitivity, meaning that the carriage states may have not been recorded with their true serotype. Therefore the classification at an observation time can sometimes be subject to error. Our approach to tackle the problem involves assuming that the observed classifications are imperfect measures of an underlying hidden disease process.

The unobserved (hidden) disease process within each pen is modelled as a multiple-state, discrete time non homogeneous Markov model. This model is an extension of the standard individual-based SIS model described in Section 3.2, but also incorporates serotype-specific information. More precisely, we define a Markov transition model with $n_s + 1$ states, in which individuals belong to a state according to their carriage status. Therefore, the possible states include being a non-carrier (state 0) or being a carrier of one of the n_s serotypes (states $1, 2, \ldots, n_s$). The model assumes that an individual can carry at most one serotype at a time: when individuals acquire a new serotype then the type is replaced by the new type. This is justified by the fact that there were only three occasions in the data set in which an individual was observed to carry different serotypes on RAMS and faecal positive samples taken on the same sampling day (see Section 2.4 for more details).

The transition rates between the any two carriage states in this Markov model, for each individual in pen p at day t, are defined for three cases:

$$h^p_{r,s}(t) = \begin{cases} \lambda^p_s(t) & r = 0, s \neq 0;\ \text{colonisation,} \\ \delta\, \lambda^p_s(t) & r, s > 0 \text{ and } r \neq s;\ \text{change of serotype,} \\ \mu_r & r \neq 0, s = 0;\ \text{clearance,} \end{cases}$$

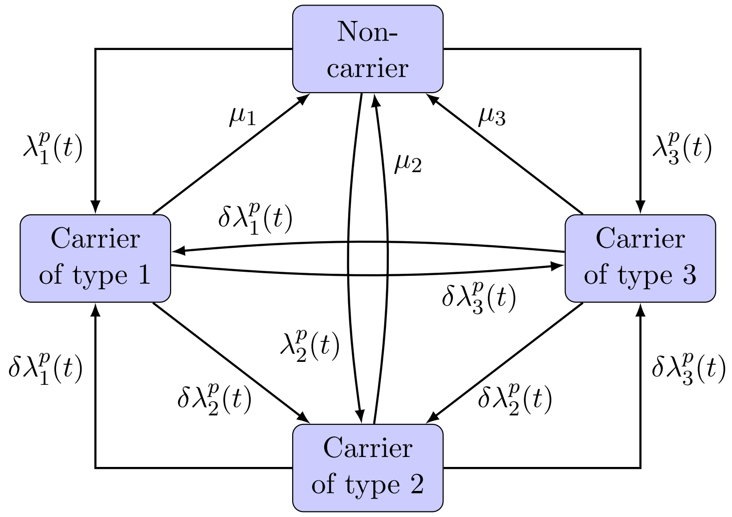
where the first case defines the colonisation rate at which a non-carrier acquires a particular serotype s at day t ($0 \mapsto s$), for which the rate depends on the serotype, the day and the individual's pen. The second case corresponds to the rate of transition from carriage of serotype r to carriage of serotype s, where $r \neq s$ ($r \mapsto s$). Between-serotype competition in colonising the host is included in the model by using an additional parameter $\delta > 0$ to scale the rate of colonisation in an individual already carrying another serotype. This parameter is assumed to be the same for all serotypes. Finally, once colonised, individuals can recover from carriage of serotype r according to serotype-dependent clearance rates μ_r that is constant over time and across different pens ($r \mapsto 0$). A simplified version of this model is presented in Figure 7.1, with only three serotypes. To fully specify the distribution of the states, we need to define a model for the initial time point. Since individuals were assigned to pens at random, we assume that at the beginning of the study each individual is colonised by serotype s independently with a serotype-dependent probability ν_s.

In addition, the model assumes that the rate at which a non-carrier individual acquires a serotype is pen-, type- and time-dependent, varying as a function of the number of other pen members carrying this particular serotype. To be more specific, for a non-carrying individual in pen p, where $p \in \mathcal{N}$ (North group) or $p \in \mathcal{S}$ (South group), the rate of colonisation of serotype s, at any given time t, is defined as the sum of two components as follows:

$$\begin{aligned}\lambda_s^p(t) &= \begin{cases}\alpha_s + \beta_s\, I_s^p(t-1), & \text{if } p \in \mathcal{S},\\[2em] \alpha_s + \gamma\,\beta_s\, I_s^p(t-1), & \text{if } p \in \mathcal{N},\end{cases}\\ &= \alpha_s + \Big(\mathbb{1}_{\{p\in\mathcal{S}\}} + \gamma\,\mathbb{1}_{\{p\in\mathcal{N}\}}\Big)\beta_s\, I_s^p(t-1), \qquad (7.1)\end{aligned}$$

where $I_s^p(t-1)$ denotes the number of carriers of serotype s in the individual's pen p at time $t-1$ and $\mathbb{1}$ denotes the indicator function. The serotype-specific terms β_s and α_s represent the rates of colonisation from contacts with other members of the pen (within-pen colonisation rate) and from sources outside of the pen (external colonisation rate), respectively. To account for differences between North (small) and South (big) pens the within-North pen colonisation rates are multiplied with γ, where γ is the relative acquisition rate in smaller versus bigger pens, as shown in Equation (7.1). This can be justified by differences in pen sizes (6m × 17m compared with 6m × 37m) and by our previous finding in Section 5.3 that animals in smaller pens are more at risk of within-pen infection. The rates of colonisation

Figure 7.1: The model graph for an individual that belongs to pen p in which, for simplicity, three serotypes are considered, denoted as 1, 2 and 3 respectively, and four carriage states. Transitions between the states are governed by rates of acquisition and clearance, as marked at each arrow. The acquisition rates depend on the number of individuals within the pen carrying that particular serotype, and for individuals already carrying another serotype the rates are adjusted by a competition parameter δ. Moreover, the rates of within-pen acquisition for individuals that belong to a small pen are scaled by a factor γ.



in a carrying individual are similarly defined, except that the rate is multiplied by a competition parameter δ, as previously described.

Regardless of the large overall number of samples in the *E. coli* dataset 1, the proportion of positive samples in the study population is fairly low, reported to be less than 10%, with only a few events per serotype. More precisely, 48 different serotypes were identified in the study population, 24 of which were found once. From the remaining 24, 7 serotypes were isolated in at least 10 RAMS and/or faecal samples, and 17 serotypes were detected in at most 5 isolates. Table 7.1 summarises the frequencies of the 7 most common *E. coli* O157:H7 PFGE serotypes in the data by pen groups, i.e. North and South.

Consequently, analysing these data using a multiple-state model where the possible states include being a carrier of one of the 48 different serotypes or being a non-carrier, presents a considerable challenge; the large number of serotype-specific parameters could lead very easily to problems in identifiability of parameters. Like most of the previous epidemic analyses, we solve this problem by dividing the serotypes into groups as follows. States of carriage are defined for the 7 serotypes most commonly recovered in this study, types A, C, G, M, O, P, T. The remaining serotypes are treated as a single group, referred as the "Unidentified" group, and

Table 7.1: Distribution of *E. coli* O157:H7 observed serotypes during follow-up of 160 cattle among 12 North and 8 South pens, July-October 2003, Canada.

Serotypes	**Observed positive samples**					
	South		**North**		**Total**	
	No.	**%**	**No.**	**%**	**No.**	**%**
A	10	3.6	22	4.6	32	4.2
C	3	1.1	7	1.5	10	1.3
G	2	0.7	14	2.9	16	2.1
M	3	1.1	7	1.5	10	1.3
O	15	5.5	17	3.5	32	4.2
P	19	6.9	0	0.0	19	2.5
T	2	0.7	28	5.8	30	4.0
Other serotypes	32	11.6	42	8.7	74	9.8
Nontyped samples	189	68.7	344	71.5	533	70.5
Total	275	100	481	100	756	100

assumed to be of the same type U. Thus their exact serotype identity is ignored. Even though this a strong assumption, it only affects a small proportion (9.8%) of the observed positive samples. For the most common serotypes we assume their own individual rates of acquisition and clearance, and the unidentified group has its own rate parameters. The disease process parameters are described in Table 7.2.

Table 7.2: Symbols and interpretations of the disease process parameters, for $s = 1, 2, \ldots, 8$.

Parameter	Interpretation
α_s	External colonisation rate for serotype s (days^{-1})
β_s	Within-pen colonisation rate for serotype s (days^{-1})
μ_s	Clearance rate for serotype s (days^{-1})
ν_s	Initial probability of carriage with serotype s
δ	Relative colonisation rate in a carrier versus non-carrier individual
γ	Relative colonisation rate in smaller versus bigger pens

Formulating the model as above substantially reduces the number of different carriage states to 9, with $n_s = 8$. To this end, we denote the carriage state of individual $c \in \{1, 2, \ldots, C\}$ in pen $p \in \{1, 2, \ldots, P\}$ on day $t \in \mathcal{T}^{c,p}$, by $X_t^{[c,p]} \in \mathcal{X}_s = \{0, 1, \ldots, n_s\}$, where $X_t^{[c,p]} = 0$ refers to the non-carriage state, state $X_t^{[c,p]} = n_s$ to carriage of the unidentified group, and state $X_t^{[c,p]} = s, 0 < s < n_s$, to carriage of one of the common serotypes. The observation (sampling) period for each individual is defined as the period from the first sample to the last one, denoted by $\mathcal{T}^{c,p} \subseteq \{1, 2, \ldots, T\}$, where the first sample in the overall study is taken at $t = 1$ and the last at $t = T$.

According to the assumptions and notation above, the model is defined as a discrete-time Markov process with time interval equal to one day, in which the current status of each individual depends on the previous status of all the individuals within the pen. The probabilities of transition between states, for any individual in pen p at time t, can be arranged in a $(n_s + 1) \times (n_s + 1)$ matrix $\mathbf{Q}^p(t)$ (time- and pen-dependent) with elements $q_{r,s}^p(t)$, for $r, s = 0, 1, 2, \ldots, n_s$ and $t \in \mathcal{T}^{c,p} \setminus \{1\}$. For convenience we start indexing the rows and columns of $\mathbf{Q}^p(t)$ from 0. The off-diagonal elements of $\mathbf{Q}^p(t)$ are specified below,

$$q^p_{r,s}(t) = \mathbb{P}\left(X^{[c,\,p]}_t = s \mid X^{[c,\,p]}_{t-1} = r, \mathbf{X}^{[-c,\,p]}_{t-1}\right) = \underbrace{\left(1 - e^{-\sum_{\substack{j=0\\j\neq r}}^{n_s} h^p_{r,j}(t)}\right)}_{\substack{\text{Probability that an}\\\text{event occurs}}} \times \underbrace{\frac{h^p_{r,s}(t)}{\sum\limits_{\substack{j=0\\j\neq r}}^{n_s} h^p_{r,j}(t)}}_{\substack{\text{Probability that}\\\text{event } r \mapsto s \text{ occurs,}\\\text{given an event occurs}}}$$

for $r \neq s$, where $\mathbf{X}^{[-c,\,p]}_{t-1}$ is the vector of the hidden states of the remaining individuals within pen p at time $t-1$. Diagonal elements in $\mathbf{Q}^p(t)$ contain the $q^p_{r,r}(t)$, which are defined as $q^p_{r,r}(t) = 1 - \sum_{\substack{j=0\\j\neq r}}^{n_s} q^p_{r,j}(t)$ so that the sum of all elements in each row equals one. Thus, using this parametrization the transition probability in the case where $r = s$ is given by $q^p_{r,r}(t) = \exp\left(-\sum_{\substack{j=0\\j\neq r}}^{n_s} h^p_{r,j}(t)\right)$, which is equal to the probability of there being no events in a Poisson process with rate $\sum_{\substack{j=0\\j\neq r}}^{n_s} h^p_{r,j}(t)$.

The observed data for an individual c in pen p are collected in prescheduled observation times, which we denote by $O^{c,p} = \left\{O^{c,p}_s \cup O^{c,p}_+\right\} \subseteq \mathcal{T}^{c,p}$, where $O^{c,p}_s$ is defined as the set of serotyped observation times and $O^{c,p}_+ = O^{c,p} \setminus O^{c,p}_s$ are the times where no serotyping was done. Moreover, let $U^{c,p} = \mathcal{T}^{c,p} \setminus O^{c,p}$ denotes the times that the individual was not tested. Let $R^{[c,\,p]}_t$ and $F^{[c,\,p]}_t$ denote the outcome of the RAMS and faecal test, respectively, recorded at time $t \in O^{c,p}$. A test result is classified as negative, denoted by 0, or positive, denoted by $+$, when $t \in O^{c,p}_+$. When a positive test is serotyped, $t \in O^{c,p}_s$, then we can further characterise the test as s-serotype positive (when a type s is detected), denoted by $s \in \{1, 2, \ldots, n_s\}$. Finally, when $t \in U^{c,p}$ the result was not reported and therefore we have a missing value denoted by “NA”.

We assume that the RAMS and faecal tests are independent conditional on the true colonisation status of the individual. Moreover, the observed states $R^{[c,\,p]}_t$ and $F^{[c,\,p]}_t$ are generated conditional on the true disease state $X^{[c,\,p]}_t$ according to a misspecification matrices $\mathbf{E}^R$ and $\mathbf{E}^F$ with elements $e^R_{r,s} = \mathbb{P}\left(R^{[c,\,p]}_t = s \mid X^{[c,\,p]}_t = r\right)$ and $e^F_{r,s} = \mathbb{P}\left(F^{[c,\,p]}_t = s \mid X^{[c,\,p]}_t = r\right)$. We distinguish two cases: tests not chosen to be serotyped and tests that were serotyped.

For the case where a positive RAMS sample was not chosen to be serotyped we have that:

$$\mathbf{E}^{R_+} = \begin{array}{c|cc|} & 0 & + \\ \hline 0 & 1 & 0 \\ 1 & 1-\theta_R & \theta_R \\ \vdots & \vdots & \vdots \\ n_s & 1-\theta_R & \theta_R \end{array} \tag{7.2}$$

and similarly, for the faecal test:

$$\mathbf{E}^{F_+} = \begin{array}{c|cc|} & 0 & + \\ \hline 0 & 1 & 0 \\ 1 & 1-\theta_F & \theta_F \\ \vdots & \vdots & \vdots \\ n_s & 1-\theta_F & \theta_F \end{array} \tag{7.3}$$

where we assume that both the RAMS and the faecal tests have 100% specificity (it is not possible to test positive when the true carriage status is non-carrier) but unknown sensitivities, denoted by $\theta_R = \mathbb{P}\left(R_t^{[c,\,p]} = + \mid X_t^{[c,\,p]} = r\right)$ and $\theta_F = \mathbb{P}\left(F_t^{[c,\,p]} = + \mid X_t^{[c,\,p]} = r\right)$, for $r = 1, 2, \ldots, n_s$.

For a positive sample that was serotyped we introduce additional parameters θ_C, θ_S and θ_U to allow for the possibility of serotype misspecification. The parameters have the following interpretations. Given that a test is found positive, θ_C denotes the probability of correctly identifying a common serotype, θ_S the probability of misclassifying a common serotype with a different common serotype, and θ_U the probability that a serotype of type U is classified as a common serotype. We assume that these probabilities are the same for both the RAMS and faecal tests. More specifically, the matrix of classification probabilities for the RAMS test is a $(n_s + 1) \times (n_s + 1)$ matrix of the form:

$$\mathbf{E}^{R_s} = \begin{array}{c|ccccccc|}
 & 0 & 1 & \cdots & \cdots & \cdots & n_s-1 & n_s\,(\text{Type U}) \\
\hline
0 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
1 & 1-\theta_R & \theta_C\,\theta_R & \dfrac{\theta_S\,\theta_R}{n_s-2} & \cdots & \cdots & \dfrac{\theta_S\,\theta_R}{n_s-2} & (1{-}\theta_C{-}\theta_S)\,\theta_R \\
\vdots & \vdots & \dfrac{\theta_S\,\theta_R}{n_s-2} & \theta_C\,\theta_R & \dfrac{\theta_S\,\theta_R}{n_s-2} & \cdots & \dfrac{\theta_S\,\theta_R}{n_s-2} & \vdots \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\
\vdots & \vdots & \dfrac{\theta_S\,\theta_R}{n_s-2} & \cdots & \dfrac{\theta_S\,\theta_R}{n_s-2} & \theta_C\,\theta_R & \dfrac{\theta_S\,\theta_R}{n_s-2} & \vdots \\
n_s-1 & \vdots & \dfrac{\theta_S\,\theta_R}{n_s-2} & \cdots & \cdots & \dfrac{\theta_S\,\theta_R}{n_s-2} & \theta_C\,\theta_R & (1{-}\theta_C{-}\theta_S)\,\theta_R \\
n_s\,(\text{Type U}) & 1-\theta_R & \dfrac{\theta_U\,\theta_R}{n_s-1} & \cdots & \cdots & \cdots & \dfrac{\theta_U\,\theta_R}{n_s-1} & (1-\theta_U)\,\theta_R
\end{array} \tag{7.4}$$

such that, for all $r \neq 0$, the probabilities $e^{R_s}_{r,0} = \mathbb{P}\left(R^{[c,\,p]}_t = 0 \mid X^{[c,\,p]}_t = r\right) = 1 - \theta_R$ and $\sum_{s=1}^{n_s} e^{R_s}_{r,s} = \theta_R$. Thus the sum of all elements in each row is equal to 1. The misclassification matrix for the faecal test is defined similarly replacing θ_R with θ_F in matrix (7.4).

Note that given the true carriage status, say r, the RAMS/faecal test results at a given observation time $t \in O^{c,p}_{+}$ are Bernoulli random variables with probabilities given by the r-th row of $\mathbf{E}^{R_+}$ and $\mathbf{E}^{F_+}$ respectively and at $t \in O^{c,p}_{s}$ are categorical random variables with probabilities given by the r-th row of $\mathbf{E}^{R_s}$ and $\mathbf{E}^{F_s}$ respectively.

7.3 Posterior sampling algorithm

Estimating the model parameters presents numerous challenges. As emphasized in the introduction, a key facet of the problem is data incompleteness; carriage states are indirectly observed through diagnostic tests and missing values are also very common, making the evaluation of the model likelihood difficult. One solution would be to marginalise over these hidden states but this would be difficult and computationally intractable because their space is high dimensional.

The approach adopted in this chapter to overcome this issue is to use Bayesian data augmentation methods, in which the unobserved carriage states are treated as additional parameters and are imputed from the data. This facilitates the use of MCMC algorithms, which are currently the most prevalent techniques for analysing data on partially observed infectious diseases and enable parameter estimation to be performed. Note that, the proposed methodology is taking into account missing data that may have occurred in between observation intervals and complete drop-outs of study individuals.

Let $\mathbf{X}_t^{[1:n_c,p]}$ be the vector of the hidden carriage states for individuals in pen p at time t and $\mathbf{X} = [\mathbf{X}_t^{[1:n_c,p]}]_{p=1,2,\ldots,P;t\in\mathcal{T}^{c,p}}$ be the whole hidden state process. Similarly, the observed longitudinal data comprises RAMS and faecal test results, denoted by $\mathbf{R} = [\mathbf{R}_t^{1:n_c,p}]_{p=1,2,\ldots,P;t\in\mathcal{T}^{c,p}}$ and $\mathbf{F} = [\mathbf{F}_t^{1:n_c,p}]_{p=1,2,\ldots,P;t\in\mathcal{T}^{c,p}}$ respectively. We use the notation $\boldsymbol{\vartheta} = (\theta_R, \theta_F, \theta_C, \theta_S, \theta_U)$ for the observation parameters, and $\boldsymbol{\phi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\nu}, \gamma, \delta)$ for the transmission parameters, where $\boldsymbol{\alpha} = [\alpha_s]_{s=1}^{n_s}$, $\boldsymbol{\beta} = [\beta_s]_{s=1}^{n_s}$, $\boldsymbol{\mu} = [\mu_s]_{s=1}^{n_s}$ and $\boldsymbol{\nu} = [\nu_s]_{s=1}^{n_s}$.

The Bayesian approach requires the specification of the prior distributions over the model parameters $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\vartheta})$, $\pi(\boldsymbol{\theta})$. We assume that prior uncertainty for these parameters can be represented by independent prior distributions. More precisely, for the serotype-specific external colonisation rates, the within-pen colonisation rates and the clearance rates, we assigned non-informative Exponential priors with means 100. The priors for δ and γ are assumed to be Exponential with rate parameter ln(2), reflecting equal prior probabilities for these parameters to be less or more than one. We also assume Beta(1,1) prior distributions for the sensitivity parameters θ_R, θ_F and θ_U. For the remaining sensitivity parameters we assume a Dirichlet prior distribution, that is, $(\theta_C, \theta_S) \sim \text{Dirichlet}(1, 1, 1)$. Finally, for the probabilities of carriage at the beginning of the study we use $\boldsymbol{\nu} \sim \text{Dirichlet}(\mathbf{1}_{n_s+1})$, where $\mathbf{1}_{n_s+1}$ is a vector with $n_s + 1$ ones.

Combining the complete data likelihood with the prior allows us to formulate the joint posterior distribution of the hidden carriage states (unobserved data) and the model parameters which can be factorised as:

$$
\begin{aligned}
\pi(\mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\vartheta} \mid \mathbf{R}, \mathbf{F}) &\propto \pi(\mathbf{R}, \mathbf{F} \mid \mathbf{X}, \boldsymbol{\vartheta})\, \pi(\mathbf{X} \mid \boldsymbol{\phi})\, \pi(\boldsymbol{\theta}) \\
&= \prod_{p=1}^{P} \prod_{c=1}^{C} \Bigg[\prod_{r=0}^{n_s} \prod_{s\,\in\,\mathcal{X}_s} \prod_{t\,\in\,O_s^{c,p}} \left[\left(e_{r,s}^{R_s} \right)^{\mathbb{1}\left\{X_t^{c,p}=r,R_t^{c,p}=s\right\}} \left(e_{r,s}^{F_s} \right)^{\mathbb{1}\left\{X_t^{c,p}=r,F_t^{c,p}=s\right\}} \right] \\
&\quad \times \prod_{r=0}^{n_s} \prod_{s\,\in\,\{0,+\}} \prod_{t\,\in\,O_+^{c,p}} \left[\left(e_{r,s}^{R_+} \right)^{\mathbb{1}\left\{X_t^{c,p}=r,R_t^{c,p}=s\right\}} \left(e_{r,s}^{F_+} \right)^{\mathbb{1}\left\{X_t^{c,p}=r,F_t^{c,p}=s\right\}} \right] \\
&\quad \times \prod_{s=0}^{n_s} \nu_s^{\mathbb{1}\left\{X_1^{c,p}=s\right\}} \times \prod_{r=0}^{n_s} \prod_{s=0}^{n_s} \prod_{t\,\in\,\mathcal{T}^{c,p}\setminus\{1\}} \left[\left(q_{r,s}^{p}(t) \right)^{\mathbb{1}\left\{X_{t-1}^{c,p}=r,X_t^{c,p}=s\right\}} \right] \Bigg] \\
&\quad \times \pi(\boldsymbol{\theta}),
\end{aligned}
\tag{7.5}
$$

where $\mathbb{1}_{\left\{X_{t-1}^{c,p}=r, X_t^{c,p}=s\right\}}$ is the indicator of individual c in pen p being in state r at time $t-1$ ($X_{t-1}^{c,p} = r$) and in state s at time t ($X_t^{c,p} = s$). The remaining indicator functions are defined similarly. The factorisation in Equation (7.5) is based on the assumption that conditionally on the model parameters, the carriage process is assumed to be independent across pens. Moreover, each individual in each pen makes an independent contribution ν_0 if uncolonised or ν_s if carrying type s at the beginning of the study ($t = 1$).

Sampling from the posterior distribution is done by constructing an MCMC algorithm that employs both Gibbs and HMC techniques. The main emphasis is on sampling the hidden carriage process $\mathbf{X}$, which was done by using a Gibbs step via the proposed iFFBS algorithm as described in Section 6.4.4. A point which is worth emphasising is that the vanilla FFBS method in our setting, with 8 serotypes and 8 animals per pen, is computationally infeasible since the transition matrix has $8^{2\times 8}$ elements. The initial probability parameters $\boldsymbol{\nu}$ and the observation parameters $\boldsymbol{\vartheta}$ are updated using Gibbs updates. The remaining parameters are updated jointly using an HMC algorithm. The algorithm requires the partial derivatives for these parameters, which can be found in Equations (F.1)–(F.4) of the Appendix F.1.

7.4 Simulation studies

In this section, we evaluate the performance of our Bayesian approach via simulation studies under different settings. In the first setting, we simulate data with the same structure as the empirical data, and in the second setting we study the sensitivity of the estimates to departures from study assumptions. In particular, we investigate the effect of the total number of samples that are serotyped.

7.4.1 Validation

A simulation study is conducted to evaluate the effectiveness of the proposed approach to estimating the model parameters. We generate data that resemble the dataset described in Section 2.4. In particular, we use the same number of North and South pens as well as the same total number of individuals per pen. RAMS and faecal samples for individual c of pen p, are collected according to the actual sampling frame, $c = 1, 2, \ldots, 8$ and $p = 1, 2, \ldots, 20$.

Both the true disease process and the observed test outcomes are obtained from the model described in Section 7.2; we use eight serotypes (seven common serotypes plus the unidentified type) and realistic parameter values, some of which are obtained from the literature. In particular, we set $\gamma = 2.4$ and $\nu_0 = 0.9$ which were the posterior medians in the inference framework of Spencer *et al.* (2015). Also, the RAMS and faecal test sensitivities are assumed to be $\theta_R = 0.77$ and $\theta_F = 0.46$, respectively. The remaining parameter values are chosen so that the serotype distribution is similar to that typical of serotypes of *E. coli*, with prevalence ranging from $2\% - 4\%$ for the unidentified group, $1\% - 3\%$ for the five most common serotypes (types A, G, O, P, T) and $1\% - 1.5\%$ for the next two most common serotypes (types C, M). We refer to this as the full model.

Note that we first generate a complete data set of carriage states. Conditional on these carriage states we then generate serotyped RAMS and faecal samples at the pre-specified observation times for each individual in the study. Once RAMS and faecal samples have been generated, we randomly choose 12 serotyped samples to remain serotyped and set the remaining as positive (+). When less than 12 positive samples occur within a pen, we select 12 by sampling the available $n < 12$ with replacement. The simulation of data is repeated 50 times. Table 7.3 summarises the simulated data as frequencies of the observed RAMS and faecal samples by pen group.

For each simulated dataset, the MCMC algorithm was run for a total of 30,000 iterations. The output was then recorded every 5^{th} iteration, after a burn-in period of 5,000 iterations. The convergence of each MCMC chain was visually assessed. The results are provided in Figure 7.2 in terms of posterior medians and 90% credible intervals. The figure illustrates that the posterior medians are generally close to the true values used to generate the simulated data. By further examining the posterior distribution of the model parameters, and in particular the 90% credible intervals, one can see that all of these distributions contain their true values, indicating that the algorithm can successfully recover parameter information.

Table 7.3: Summary of our simulated data. The numbers of observations are based on 50 simulated data sets, each having the same structure and sampling times as in the real data. The model consists of 7 common serotypes and the unidentified group (type U), with serotype-specific colonisation and clearance rates.

Serotypes	**Number of observed samples**[†] **(min, max)**[‡]		
	South	**North**	**Total**
Type A	9 (2, 24)	16 (6, 27)	26 (13, 41)
Type C	6 (0, 18)	7 (0, 24)	12 (8, 27)
Type G	8 (0, 21)	16 (5, 36)	24 (13, 45)
Type M	5 (0, 13)	7 (1, 19)	14 (8, 20)
Type O	10 (1, 20)	16 (3, 26)	26 (13, 37)
Type P	8 (0, 21)	15 (5, 29)	25 (11, 37)
Type T	7 (0, 23)	14 (6, 26)	21 (11, 32)
Type U	31 (18, 52)	47 (31, 72)	81 (51, 101)
Non-typable	77 (32, 143)	450 (291, 629)	524 (341, 673)
Negative swabs	3379 (3308, 3434)	4717 (4537, 4889)	8101 (7947, 8296)
Serotyped pairs[§]			27 (17, 36)
Contradictory pairs[¶]			6 (1, 12)

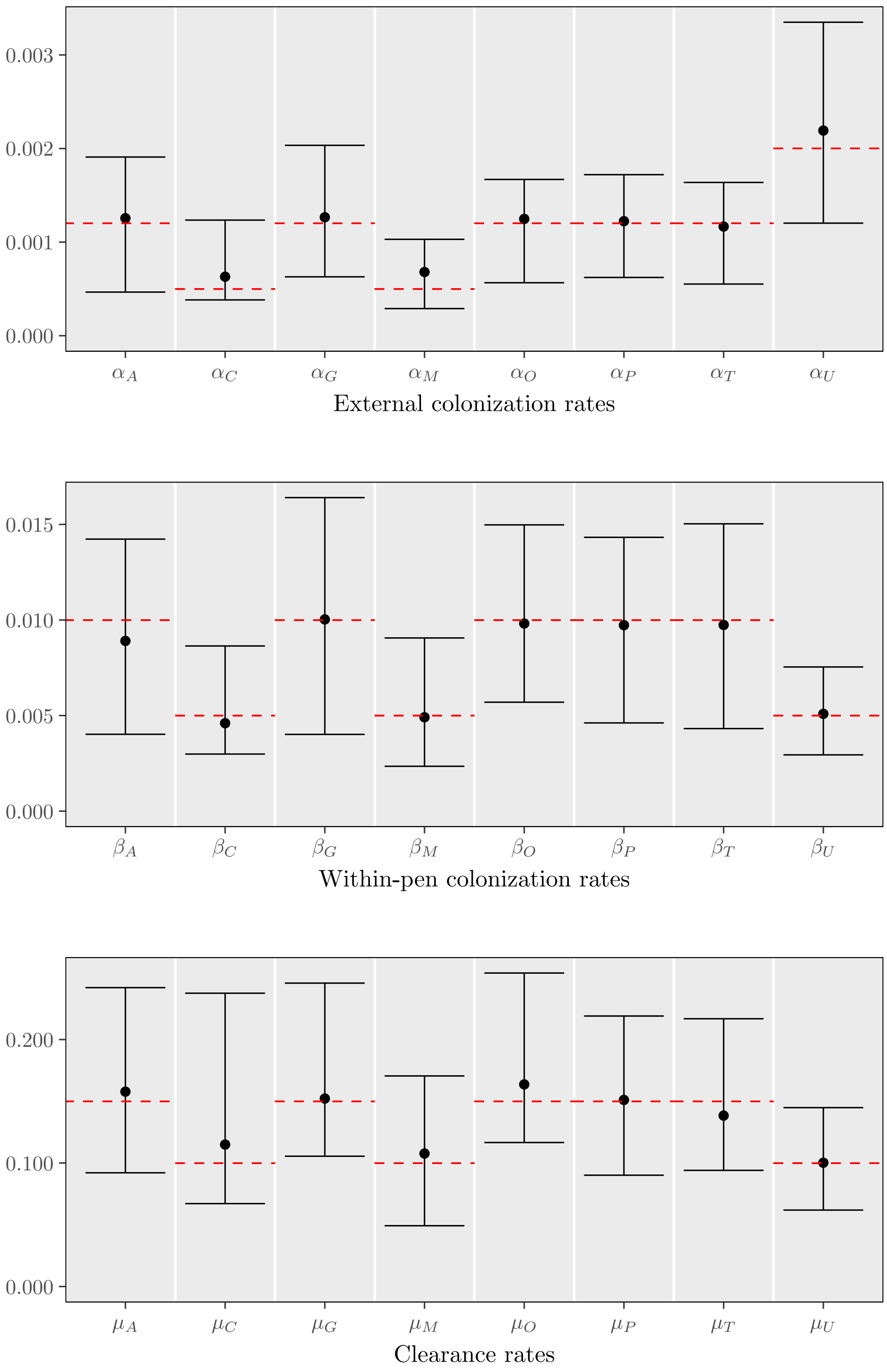
[†] The median number of samples in the 50 simulated datasets.

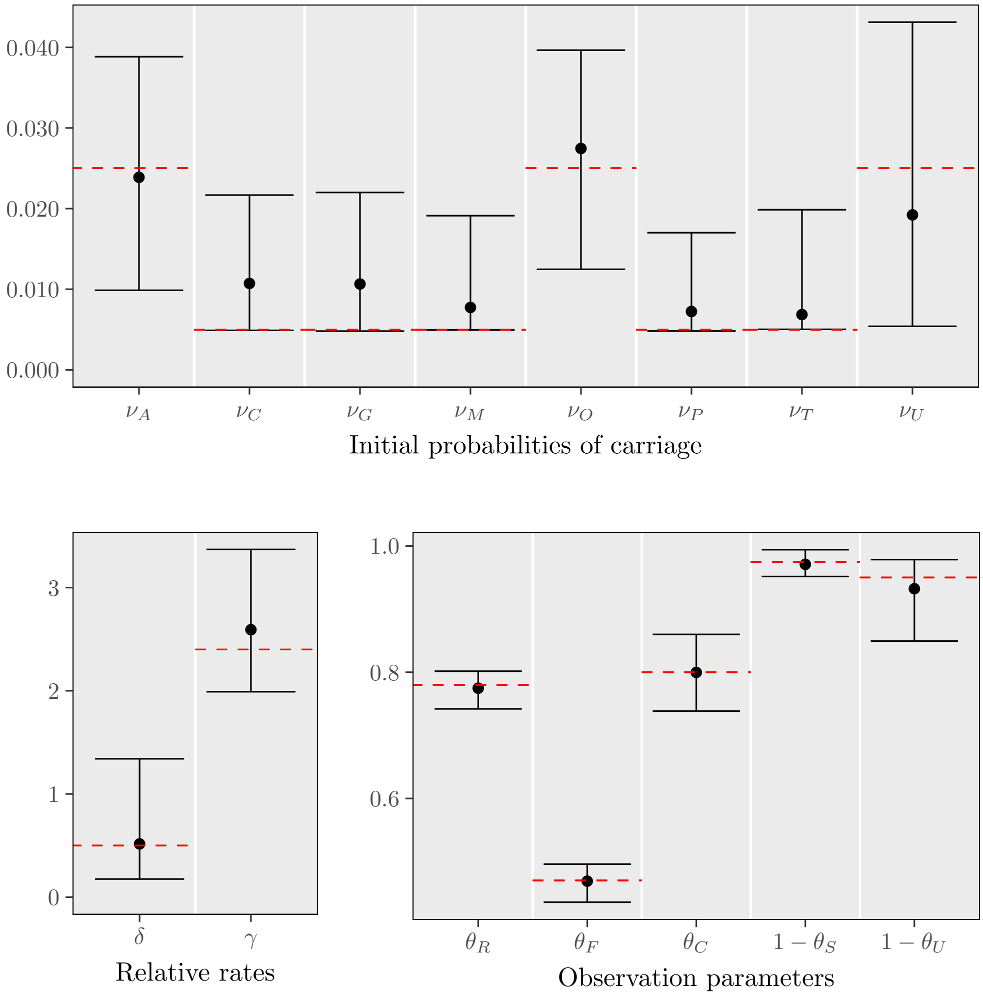
[‡] The minimum and maximum number of samples in the 50 simulated datasets.

[§] When a pair of positive RAMS and faecal samples are chosen to be serotyped. We define as pair samples that were taken on the same sampling day, from the same individual.

[¶] Number of serotyped pairs of different serotypes.

FIGURE 7.2: Marginal posterior summaries over 50 simulated data. Dots denote the posterior median and error bars indicate the 90% quantile intervals of the 50 posterior medians. Dashed red lines indicate the true value used to generate the simulated data.





Another important task in the estimation procedure is recovering the hidden carriage process. Using the augmented states of carriage in each MCMC iteration, one can estimate the probability that an individual is colonised by a specific serotype or not colonised for every day in the study, regardless of whether RAMS and faecal tests were taken on that day. To access the accuracy of the method we compare the posterior probability of colonisation for each individual of a given pen over the sampling period with the actual carriage states. Results for 3 pens are shown in Figures 7.3(a)-(c). The plot for each individual is divided into 3 panels. In the bottom panel we show the true disease status for each day in the study. The middle panel contains the posterior probabilities of colonisation. Finally, in the top panel we include the test results which are imperfect measures of the true underlying process in the bottom panel. Note that samples are taken twice per week; line 1 refers to the RAMS samples whereas line 2 refers to the faecal samples.

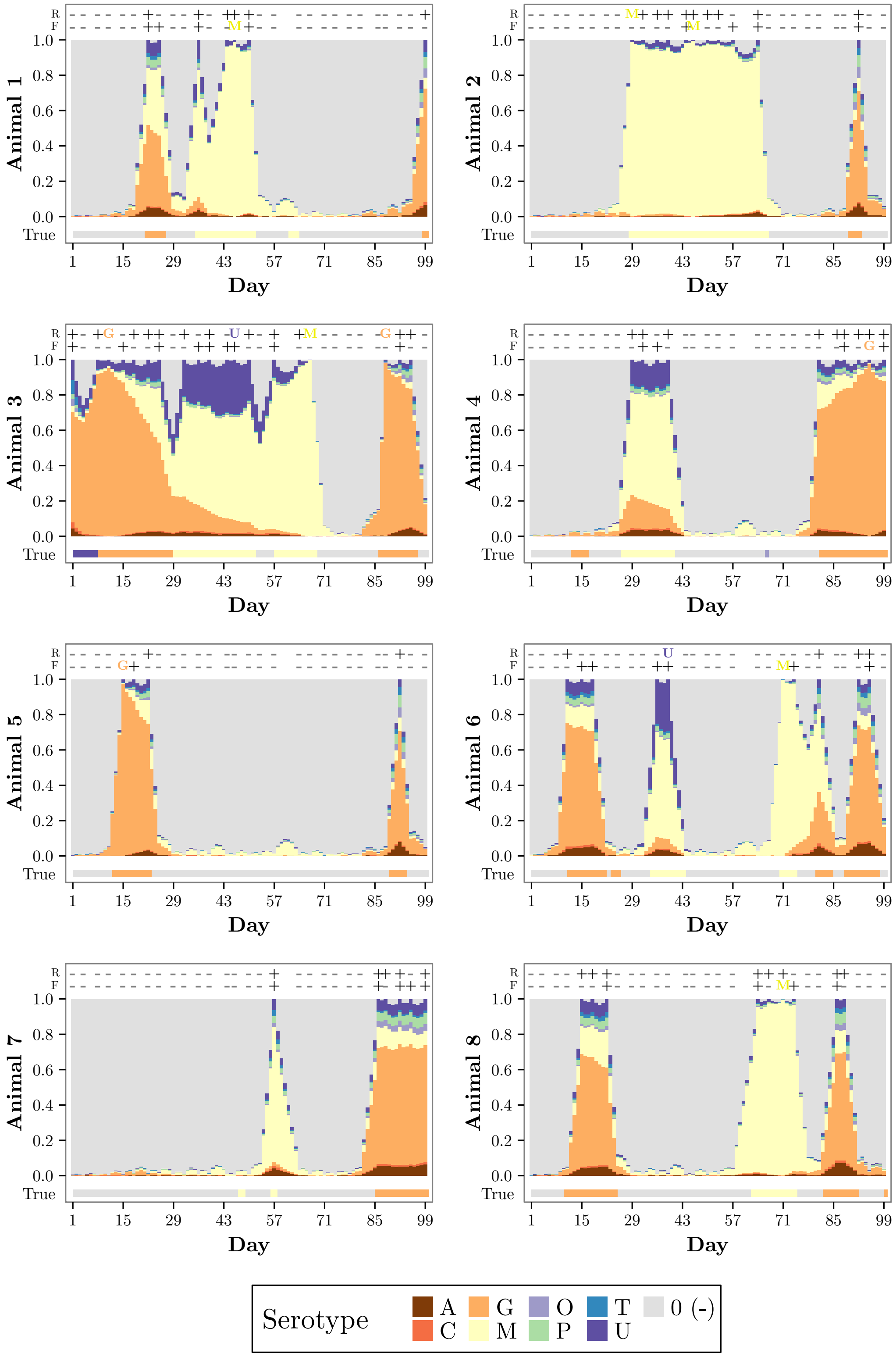
We generally see that for most of the individuals the predicted serotypes (the ones that have the highest posterior probability) match the true carriage status in the simulated data. This indicates that the method correctly reproduced the unobserved disease process. As expected, in all figures the posterior probability of colonisation by any serotype sums to 1 when the animal is tested positive by RAMS and/or faecal tests. This is because we have assumed that the tests have 100% specificity. On the other hand, when both test results are negative, the posterior probability of colonisation by some serotype can be any value between and including 0 and 1. Moreover, when sequences of positive results separated by negative results occur, it is difficult for the method to distinguish between false negative results and re-acquisition; this can be seen from the grey spikes that appear between positive results.

Additionally, our method can predict the serotype of a positive individual even though their samples are not serotyped. For example, the samples of individual 7 in Figure 7.3(a) are never serotyped but the method correctly predicts the type. This can happen even if the test indicates the wrong type; for individual 6 at day 39 the RAMS test found type U but the method correctly predicts the individual as type M, borrowing information from the rest of the individuals within the pen. However, it is also possible that a serotype is misclassified, see for example individual 3 at the beginning of the study. For this individual, the closest recorded serotypes appear on day 11 and so the method assigns high probability to the same type as in day 11. Further, when no strong information is available, we see that the probabilities of the different serotypes are roughly equal to their average relative prevalences. This is the case for animals 2 and 6 in Figure 7.3(b) at the end of the study.

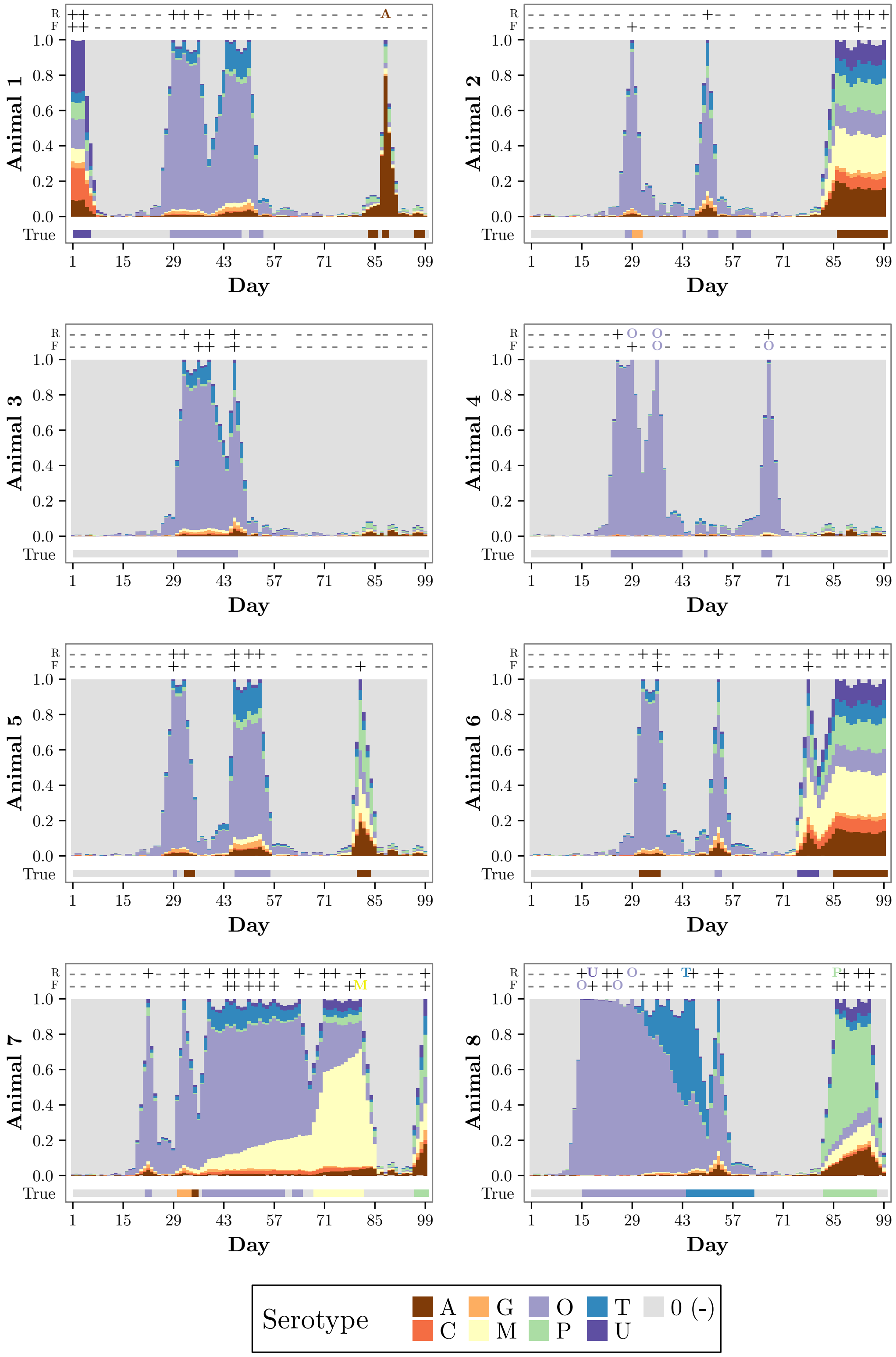
For individual 5 in Figure 7.3(a), we see that the method assigns a non zero probability of the individual being colonised by serotype M during the period around day 60 when no samples were taken. This happens because a few days earlier and after there were individuals that had an M positive result (individuals 1, 2, 3, 6 and 8) and so the method allows for the possibility that individual 5 was also colonised. Finally, it is worth noting that it is possible for short periods of carriage to remain completely unobserved. As an example, see individual 2 in Figure 7.3(c) before day 15.

We further evaluate the performance of our method for imputing the hidden carriage states by plotting the ROC curve for each serotype and calculating the area under curve. The ROC curve is a plot of the true-positive rates (the proportion of

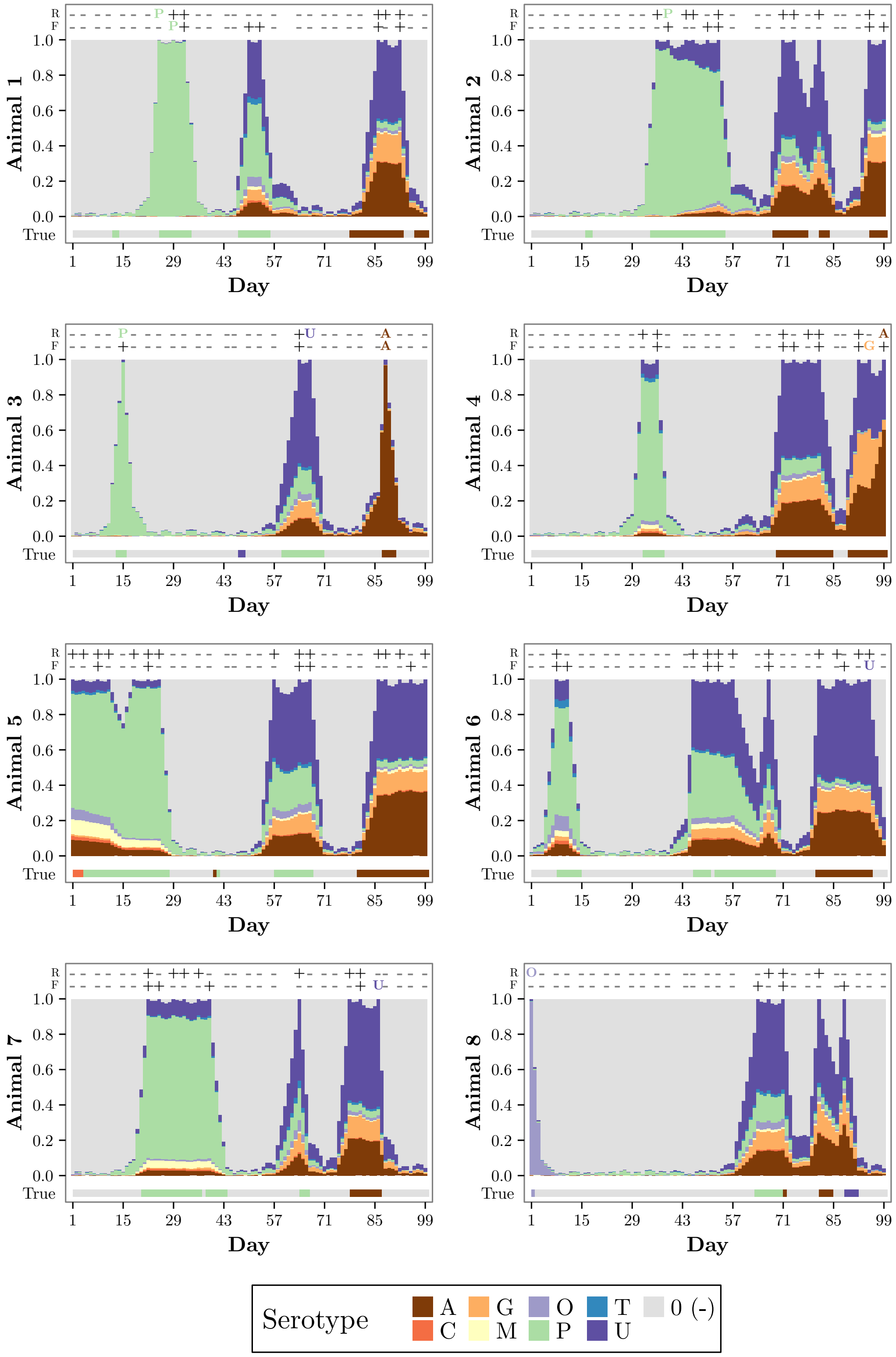
Figure 7.3: Posterior probability of colonisation over time with separate plots for each individual within a pen. In each figure the top panel contains the observed test results, where the first line represents the outcome of RAMS samples and the second line represents the outcome of faecal samples. Samples are taken twice per week as in the real dataset. "-" indicates negative sample, "+" indicates that the sample was positive but not chosen for serotyping; otherwise, serotype name is given. The bottom panel shows the true carriage status of each individual per day.



(a) Simulated data 1.



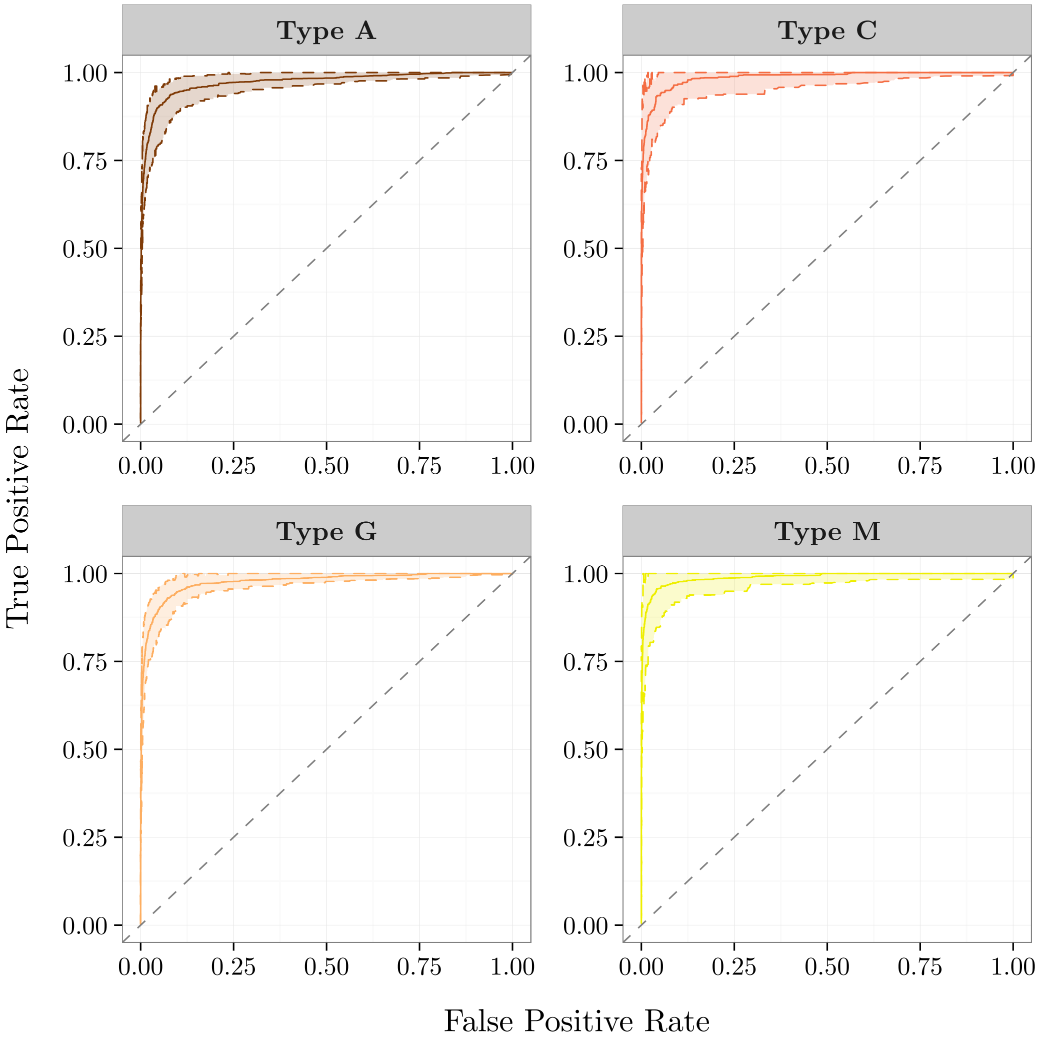
(b) Simulated data 2.

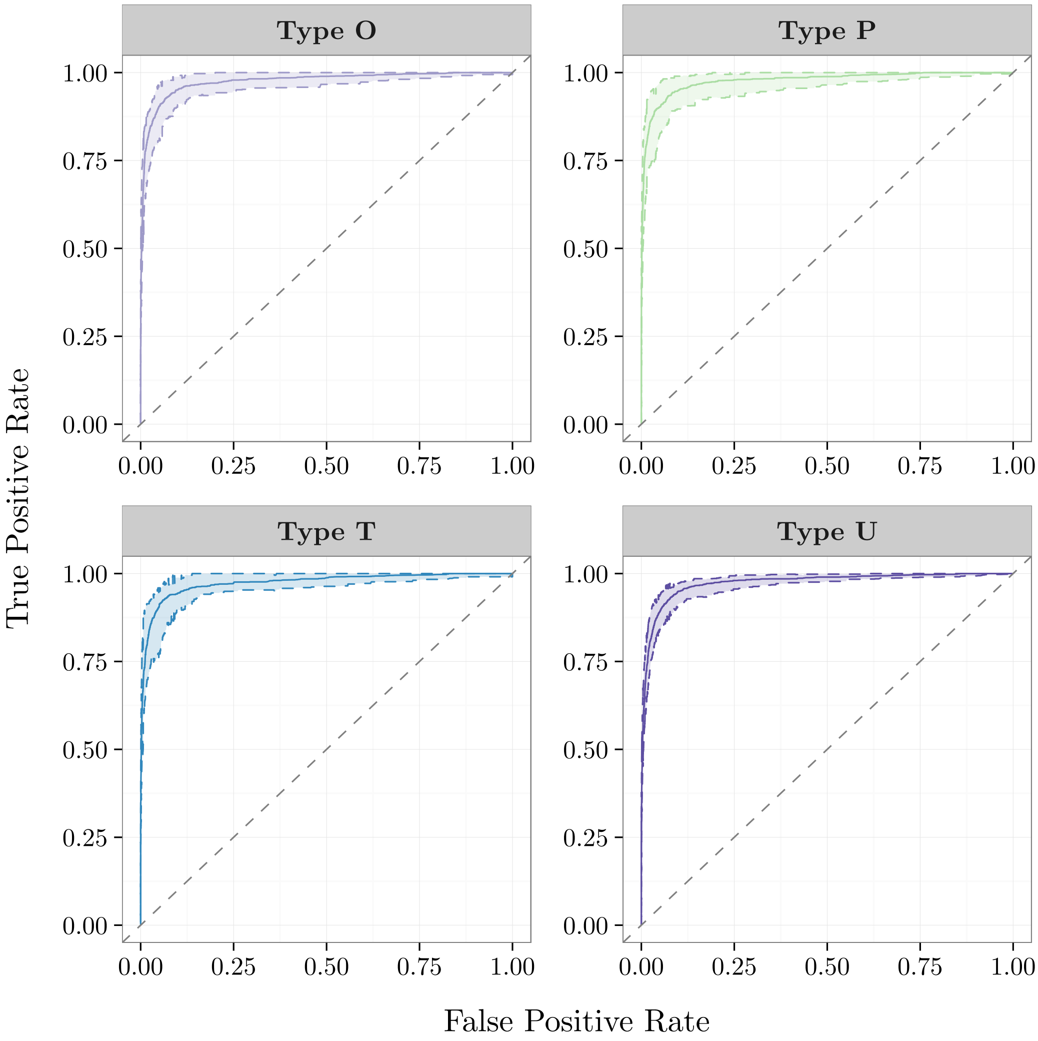


(c) Simulated data 3.

correctly detected s-serotype carriages) against false-positive rates, calculated over a range of possible threshold values on the posterior colonisation probabilities. We count a s-serotype carriage as detected if the posterior probability of colonisation by serotype s is greater than a given threshold. In Figure 7.4 we plot, for each serotype, the ROC curves produced from 50 simulated data sets. In all serotypes, the ROC curve is located close to the top left corner and the median AUC value is found to be above 0.95, indicating that our method successfully reproduces the incidence of colonisation.

FIGURE 7.4: Receiver Operating Characteristic (ROC) curves for each serotype. The solid lines correspond to the median over the 50 simulated data sets and the dashed lines indicate the 90% quantile intervals of the 50 posterior medians.





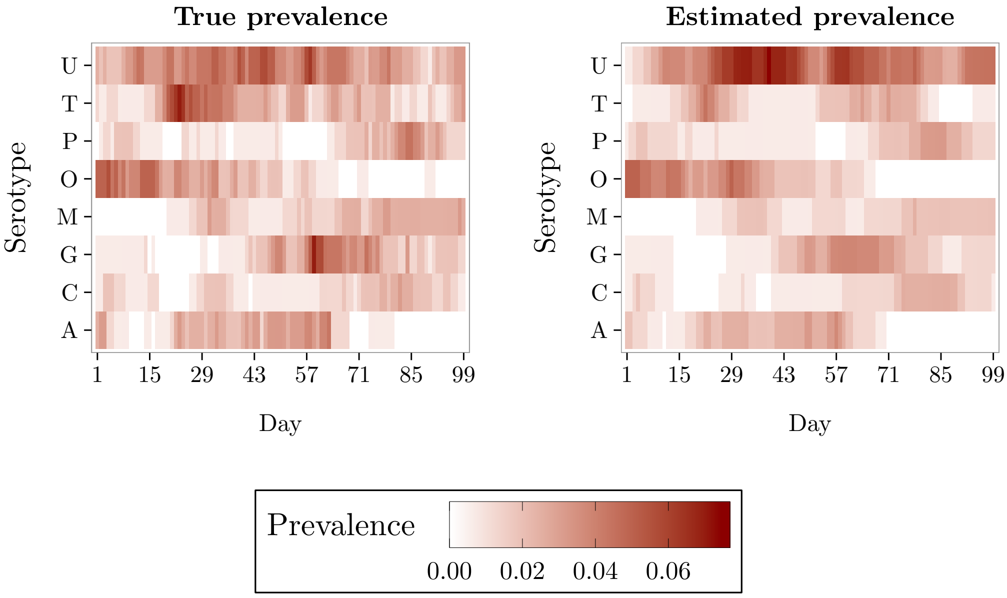
7.4.2 Robustness to model misspecification

In this section we apply our method to data that were simulated under a different epidemiological scenario in order to access its robustness to model misspecification. In particular, we assume that the transitions from carriage of one serotype to another have occurred through the non-carriage state: carriage needs to be cleared before colonisation by another serotype. This is achieved by generating data in which the competition parameter δ is 0 and the remaining parameters are as in Section 7.4.1. A full model is then estimated.

The estimated posterior prevalence of colonisation for each serotype over the sampling period are provided in Figure 7.5 and show that the model correctly reproduces the transmission dynamics of each serotype during the study period.

Further details on the posterior of the model parameters over 50 simulated datasets are presented in Figure F.1 in the Appendix and reveal that the estimates are close to the true values. We also find that the posterior median of the competition parameter is 0.18 with 90% credible interval [0.007, 0.682]. Since δ is constrained to be positive in our estimation algorithm, its 90% credible interval does not contain 0, however it is close to zero.

Figure 7.5: Performance assessment of our method. The true prevalence of each serotype (left) is compared to the predicted prevalence using the full model. Data are generated with $\delta = 0$.



7.4.3 Sensitivity analysis: total number of serotyped samples

In what follows, we investigate the performance of our approach subject to the amount of serotyped samples per pen. In particular, we first simulate a complete data set of fully serotyped test results for 8 serotypes as in Section 7.4.1. Of these data, we create 4 datasets by randomly selecting 5, 10, 15 and 20 of the serotyped observations per pen to remain serotyped, setting the rest as untyped positives and then fit our model. The 4 datasets are obtained from the same underlying carriage states and reflect situations with sparse, moderate and dense serotyping. To avoid sampling biases, we repeat the randomisation of the 4 datasets 50 times. For reference, we also fit our model to the complete data where all positive tests have been serotyped.

The results are summarised in Figure 7.6. We report the median of the posterior median estimates along with the median of the upper and lower limits of the 90% credible intervals, as obtained from the 50 randomisations. As expected, the performance of our method depends on the amount of observed serotypes. More specifically, as the total number of serotypes increases, the accuracy of the estimate increases for all model parameters. However, even when the serotyping is very sparse we see that our method performs fairly well. In particular we see that the true parameter values are included within the intervals provided across all scenarios considered. Nevertheless, changing the number of serotyped samples from 5 to 10 leads to a large improvement in precision of the estimates.

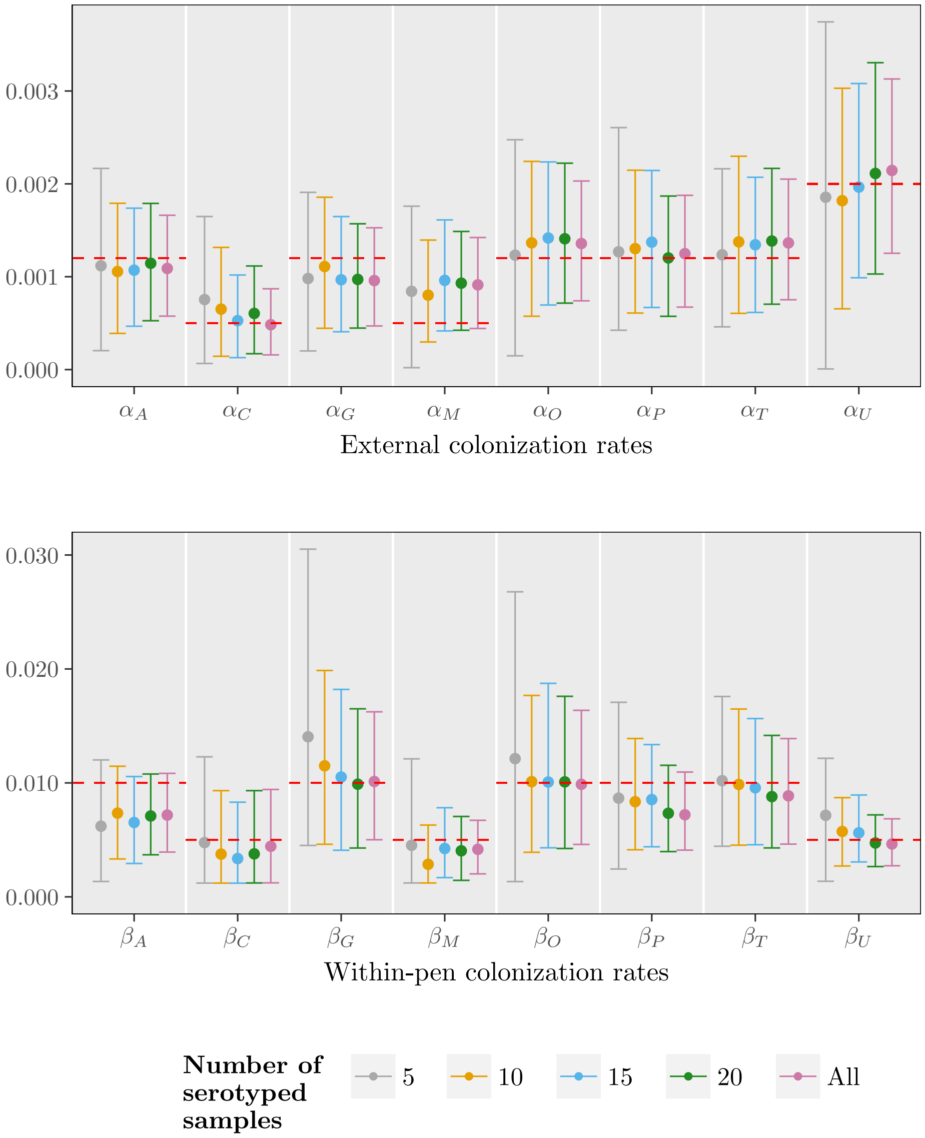
In Figure 7.7, we additionally provide ROC curves (median over 50 randomizations) to examine how well our method recovers the underlying carriage states with varying levels of serotyping. Dense serotyping is associated with higher performance; however the performance remains high even when we only observe 5 serotypes per pen. Finally, in Table 7.4 we report the median estimated number of type-specific transitions in the carriage process as estimated from 50 MCMC runs. For comparison we also report the true number of transitions in the complete data. We see that the model adequately fits the data, except maybe for the transitions from 0 to 0 and from r to r which are slightly underestimated and overestimated, respectively. However, all of these quantities lie within the corresponding 90% credible intervals.

As a general remark, taking more than 10 serotyped samples was sufficient for accurate estimation of the parameters and the hidden disease process. We therefore conclude that the amount of serotype information available for our real data application is satisfactory. Nevertheless, it is important to study through simulation the appropriate number of serotypes in applications with longer sampling intervals.

Table 7.4: Medians of the number of type-specific transitions over 50 simulated data sets with varying levels of serotyping.

Transitions	**Number of serotyped samples per pen**					**True**
	5	10	15	20	All	
$r \mapsto s \ (r \neq s)$	22	18	21	26	26	19
$0 \mapsto s \ (s \neq 0)$	289	284	284	286	289	276
$r \mapsto 0 \ (r \neq 0)$	279	274	274	275	279	265
$r \mapsto r \ (r \neq 0)$	2002	1998	1992	1990	1966	1917
$0 \mapsto 0$	12893	12911	12914	12920	12932	13018

Figure 7.6: Marginal posterior summaries of the model parameters over 50 simulated data sets for different numbers of serotyped samples per pen: grey for 5, yellow for 10, blue for 15, green for 20 and pink for the full pen. Dots denote the posterior median and error bars indicate the 90% credible intervals. Dashed red lines indicate the true value used to generate the simulated data.



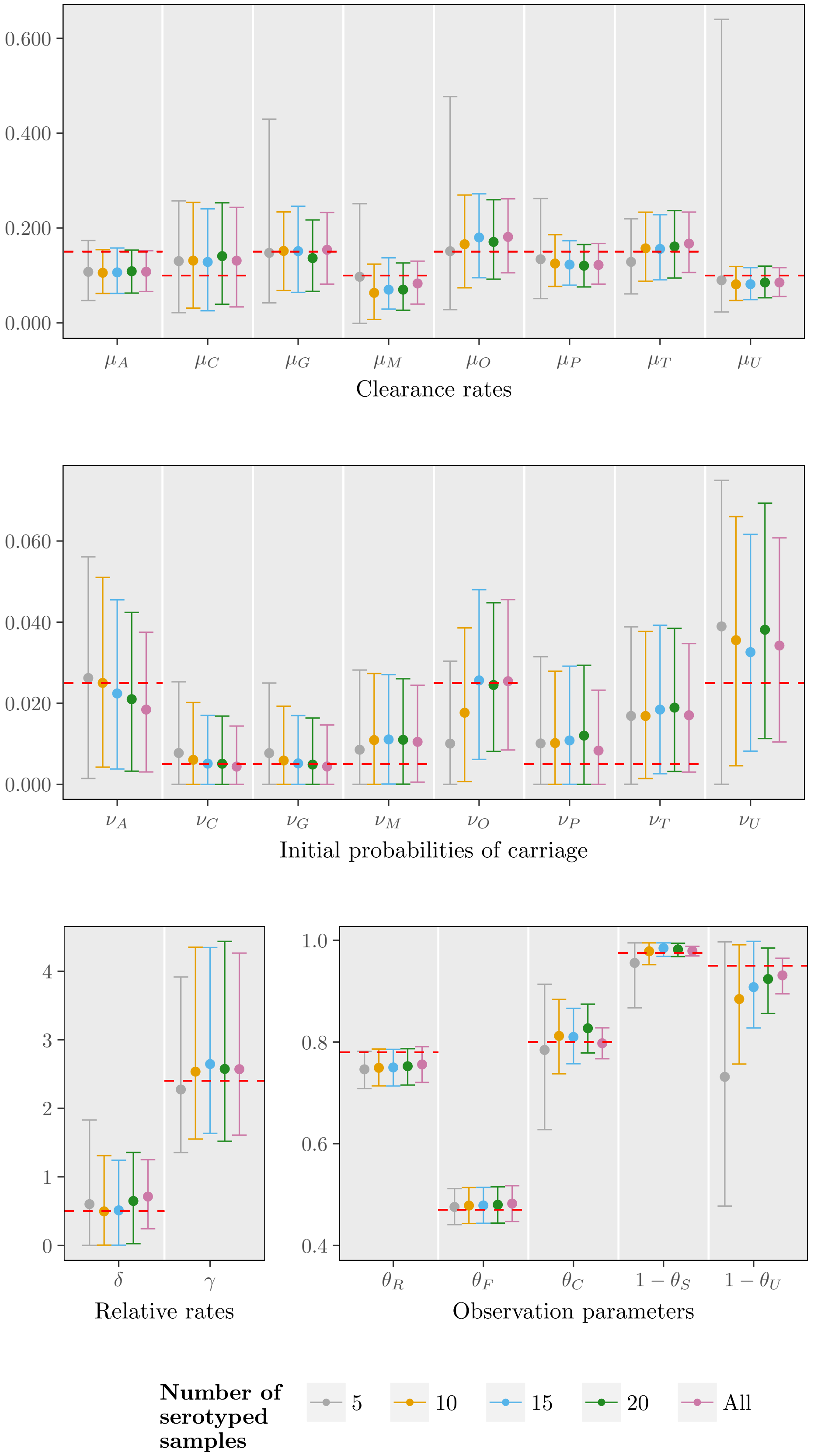
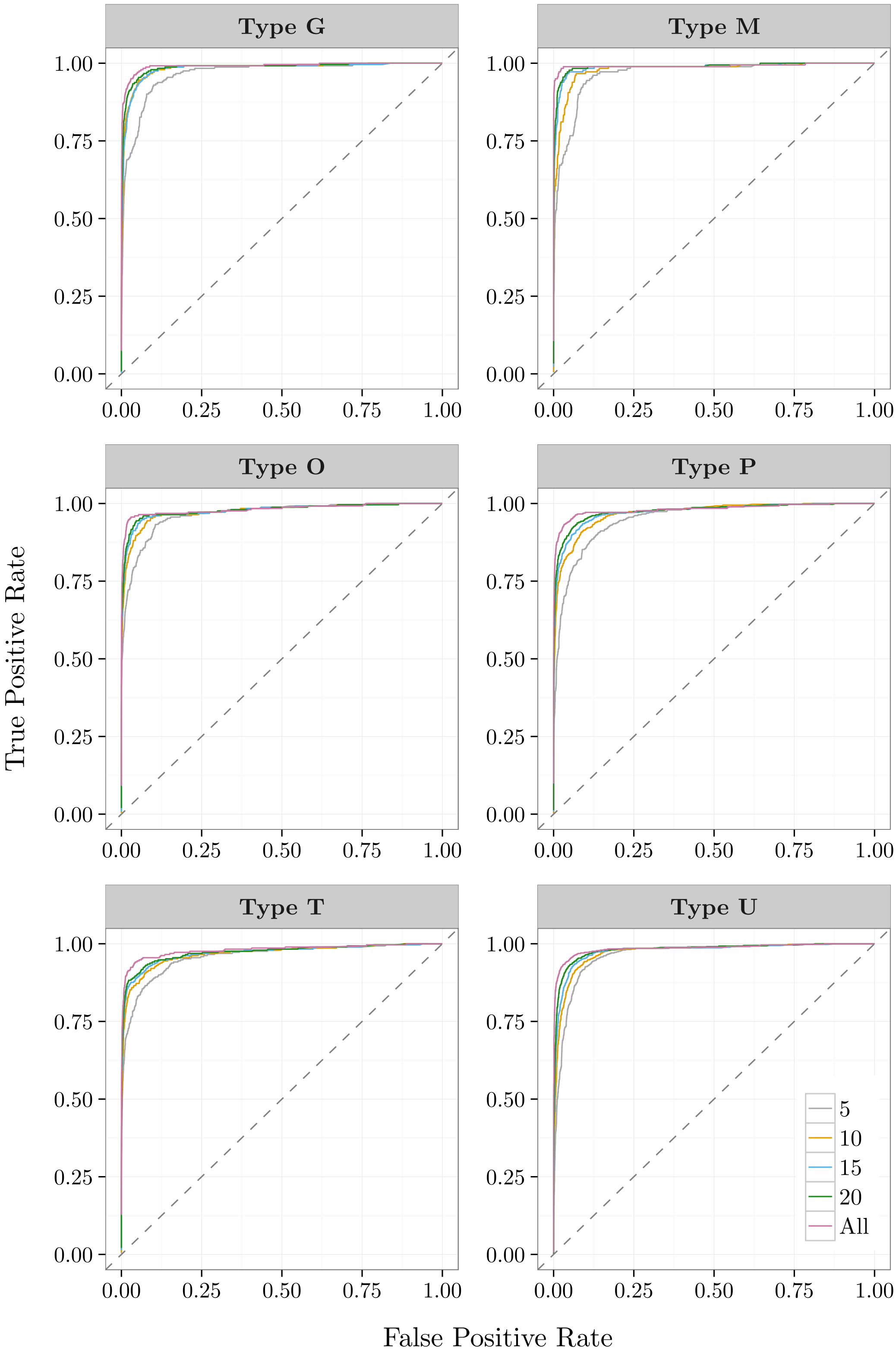
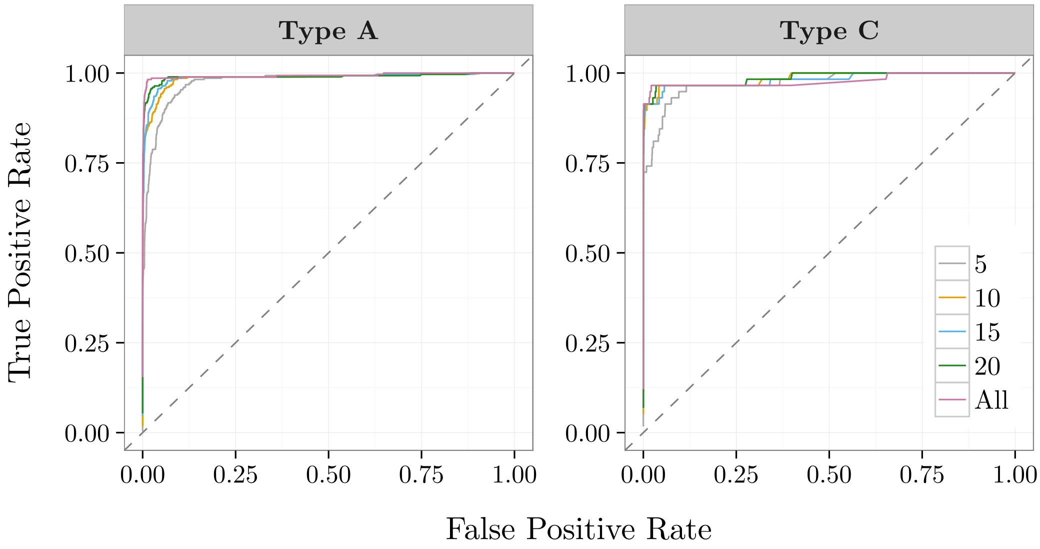

0.600
0.400
0.200
0.000
μ_A
μ_C
μ_G
μ_M
μ_O
μ_P
μ_T
μ_U
Clearance rates
0.060
0.040
0.020
0.000
ν_A
ν_C
ν_G
ν_M
ν_O
ν_P
ν_T
ν_U
Initial probabilities of carriage
4
3
2
1
0
δ
γ
Relative rates
1.0
0.8
0.6
0.4
θ_R
θ_F
θ_C
$1-\theta_S$
$1-\theta_U$
Observation parameters
Number of serotyped samples
5
10
15
20
All

Figure 7.7: Receiver Operating Characteristic (ROC) curves for each serotype. The solid lines correspond to the median over the 50 simulated data sets with varying levels of serotyping.





7.5 Real Data Analysis

In this section, we apply our Bayesian data augmentation approach to the observed *E. coli* 0157:H7 data described in Section 2.4. Our goal is to obtain estimates for the epidemiologically important parameters and also investigate possible differences between serotypes in carriage colonisation and clearance. Therefore, we fit the full model, as described in Section 7.2, to the data.

We run the MCMC for 30,000 iterations, discarding the first 5,000 as a burn-in and save every 5 iterations to obtain 5,000 samples from the posterior. We use the same priors as in the simulation studies of Section 7.4; nevertheless, we also perform a sensitivity analysis in which we use alternative priors. Convergence is accessed by visual inspection of posterior trace plots for all 39 model parameters, shown in Figure F.3 of the Appendix. We also checked that estimates were robust to a change in the initial values. Convergence of the hidden state process is also visually assessed. For example, estimated posterior probabilities of colonisation are shown for individuals in pens 3 and 7 and can be found in Figures 7.8 and 7.9 respectively. Results are summarised below.

Tables 7.5 shows the posterior median estimates of the transmission parameters, along with 90% credible intervals. We see that the external colonisation rate for the unidentified group, type U, is uniformly higher than the rest of the types. This is due to the fact that α_U accounts for all acquisitions of the 41 serotypes in the group. The lowest external colonisation rate belongs to serotype P, following by M, C and G. However, most of the differences are not significant. This is suggested

Figure 7.8: Posterior probability of colonisation over time with separate plots for each individual within Pen 3 in the *E. coli* data. In each figure the top panel contains the observed test results, where the first line represents the outcome of RAMS samples and the second line represents the outcome of faecal samples. Samples are taken twice per week; "-" indicates negative sample, "+" indicates that the sample was positive but not chosen for serotyping, otherwise, serotype name is given.

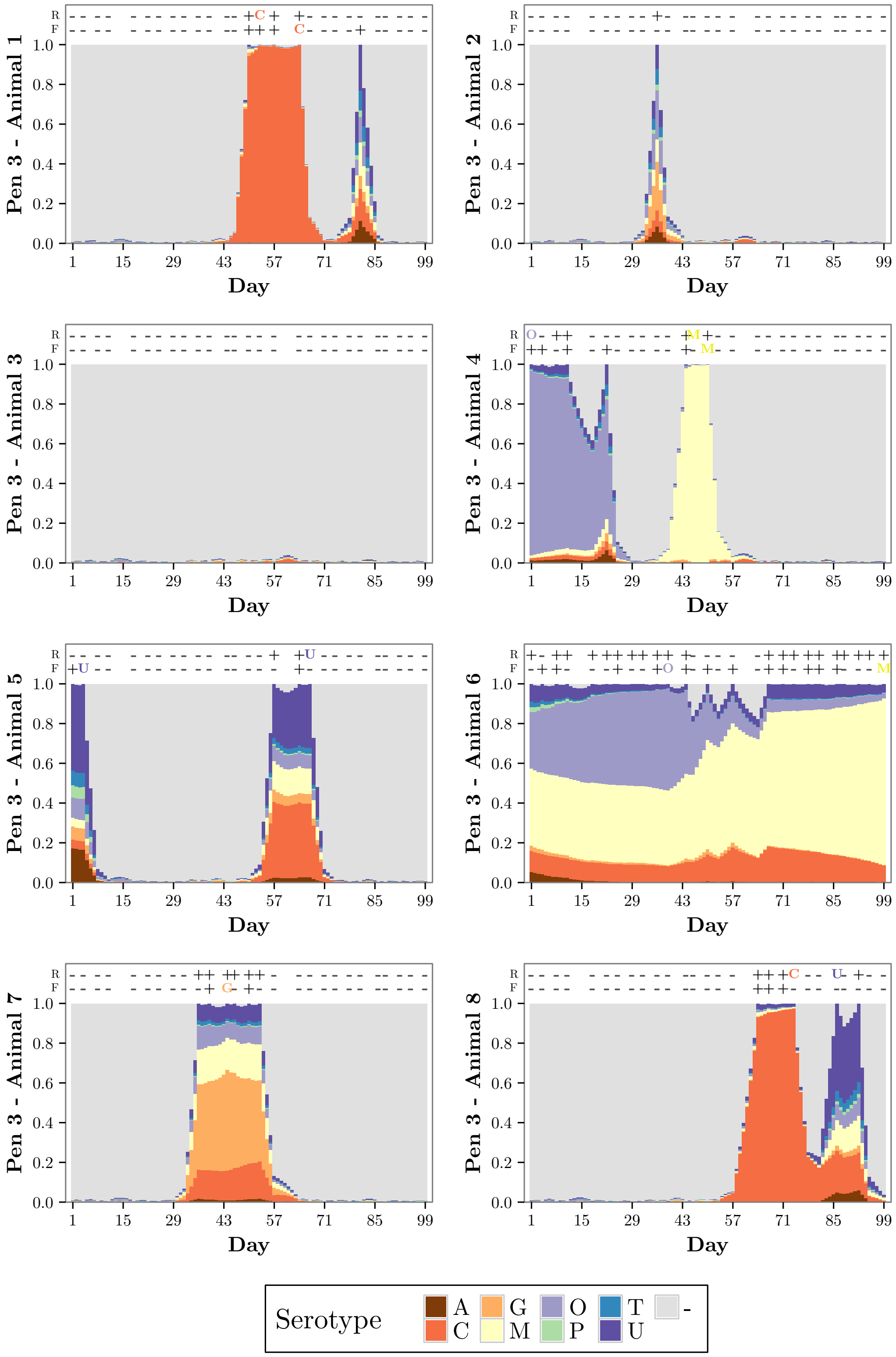
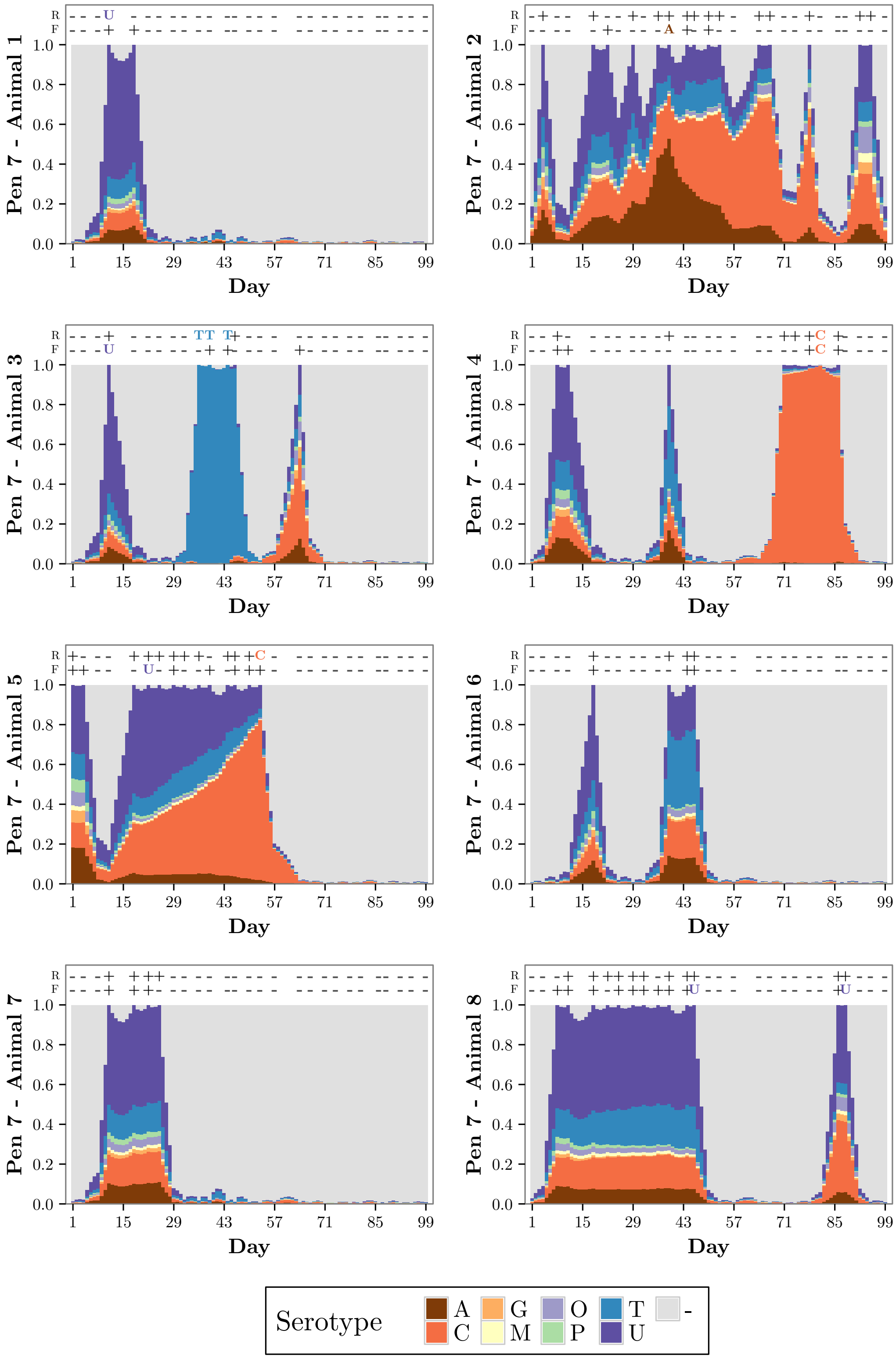


Figure 7.9: Posterior probability of colonisation over time with separate plots for each individual within Pen 7 in the *E. coli* data. In each figure the top panel contains the observed test results, where the first line represents the outcome of RAMS samples and the second line represents the outcome of faecal samples. Samples are taken twice per week; "-" indicates negative sample, "+" indicates that the sample was positive but not chosen for serotyping, otherwise, serotype name is given.



by the overlap of the credible intervals for all parameters except α_U and α_P. In the top panel of Figure 7.10 we show the mean standardised posterior differences between all pairs of serotypes:

$$\frac{\bar{D}_{\alpha_{kr}}}{\sqrt{\frac{\sum_{j=1}^{J}\left(\alpha_k^{(j)} - \alpha_r^{(j)} - \bar{D}_{\alpha_{kr}}\right)}{J-1}}}, \quad \bar{D}_{\alpha_{kr}} = \frac{1}{J}\sum_{j=1}^{J}\left(\alpha_k^{(j)} - \alpha_r^{(j)}\right)$$

where $\alpha_i^{(j)}$ is the j-th posterior draw for the serotype-specific parameter α_i for $i = 1, 2, \ldots, n_s$, and J is the total number of MCMC iterations. High absolute values in the image indicate strong difference between serotypes. The figure confirms our previous observations. Moreover, we see that the sum of the 8 serotype-specific external colonisation rates ($\sum_{s=1}^{n_s} \alpha_s = 0.008$) derived here is in close agreement with the external colonisation rate estimated in our non-serotype specific analysis in Section 3.6.1.1 ($\alpha = 0.009$). In particular, due to the assumption of different rates per serotype, the overall external force of colonisation divides between the serotypes that are analysed.

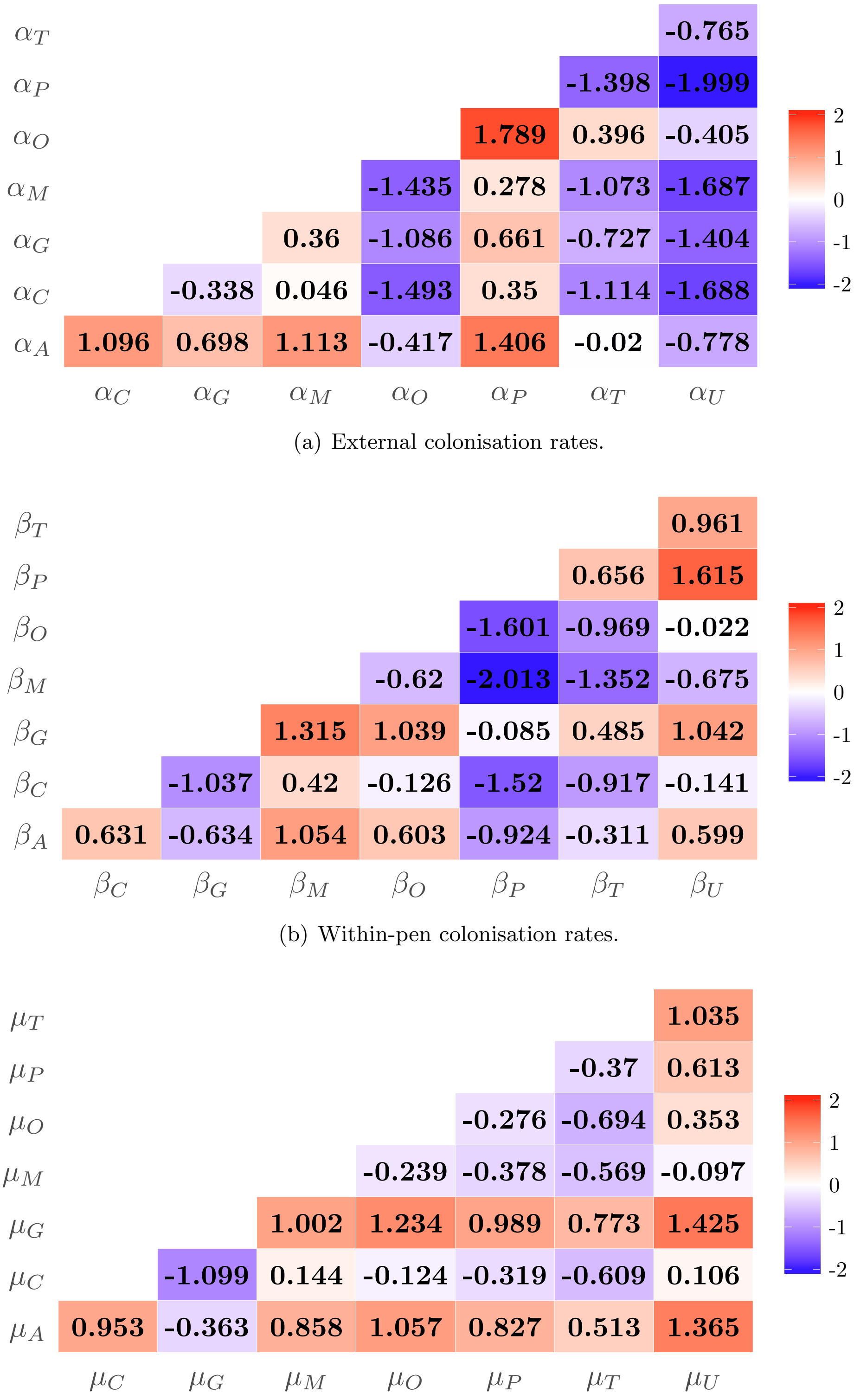
The posterior median for the within-pen colonisation rate is almost 4 times higher for serotype P (0.016 per day) compared to serotype M (0.004 per day) with non-overlapping 90% credible intervals, which suggests that there are differences in the within-pen colonisation rates between the studied serotypes. The estimates of within-pen colonisation rate for remaining serotypes are between these two values. Mean standardised posterior differences are shown in the middle panel of Figure 7.10.

Results suggest no significant differences between durations (1/clearance rate) of carriage of serotypes. In particular, in Table 7.5 we see that all parameters have overlapping credible intervals. Also, in the bottom panel of Figure 7.10 we observe a maximum mean standardised posterior difference of 1.42, which proves our claim.

As a general finding, we observe that serotype P appears to have the highest within-pen but the lowest external colonisation rate suggesting that it is mainly transmitted through contact between animals in the same pen. Similarities between colonisation rates are found for serotypes A with T and also M with C. The latter serotypes (M and C) are the less prevalent (Figure F.2 in Appendix, as calculated from the latent carriage process) which explains their low within-pen and external transmission rates.

The relative colonisation rate in smaller versus bigger pens is estimated as 1.659 with 90% credible interval being [1.142, 2.490]. The interval does not contain 1

Figure 7.10: Mean standardised posterior differences between all pairs of serotypes in the *E. coli* data.



	α_C	α_G	α_M	α_O	α_P	α_T	α_U
α_T							-0.765
α_P						-1.398	-1.999
α_O					1.789	0.396	-0.405
α_M				-1.435	0.278	-1.073	-1.687
α_G			0.36	-1.086	0.661	-0.727	-1.404
α_C		-0.338	0.046	-1.493	0.35	-1.114	-1.688
α_A	1.096	0.698	1.113	-0.417	1.406	-0.02	-0.778

(a) External colonisation rates.

	β_C	β_G	β_M	β_O	β_P	β_T	β_U
β_T							0.961
β_P						0.656	1.615
β_O					-1.601	-0.969	-0.022
β_M				-0.62	-2.013	-1.352	-0.675
β_G			1.315	1.039	-0.085	0.485	1.042
β_C		-1.037	0.42	-0.126	-1.52	-0.917	-0.141
β_A	0.631	-0.634	1.054	0.603	-0.924	-0.311	0.599

(b) Within-pen colonisation rates.

	μ_C	μ_G	μ_M	μ_O	μ_P	μ_T	μ_U
μ_T							1.035
μ_P						-0.37	0.613
μ_O					-0.276	-0.694	0.353
μ_M				-0.239	-0.378	-0.569	-0.097
μ_G			1.002	1.234	0.989	0.773	1.425
μ_C		-1.099	0.144	-0.124	-0.319	-0.609	0.106
μ_A	0.953	-0.363	0.858	1.057	0.827	0.513	1.365

(c) Clearance rates

which demonstrates a small but significant difference between North and South pens. Our finding is in agreement to Spencer *et al.* (2015) who obtain a larger estimate of the relative colonisation rate γ. However their analysis does not account for serotype information. The median relative colonisation rate in a carrier versus non-carrier individual is 0.523 (90% CI [0.005, 1.201]) which indicates that individuals colonised by a serotype are less likely to be infected by another type.

Table 7.6 show posterior summaries for the observation parameters. As in previous analysis of this dataset in Section 3.6.1, we find that the test sensitivities θ_R and θ_F are 0.76 and 0.46, respectively. The model estimates that 78.5% of the observed common serotypes are correctly classified as the right type, 3.5% are misclassified as another common type and the remaining 18% are misclassified as type U. Finally, we estimate that 95.5% of the observed U serotypes are correctly classified as U.

Table 7.5: Estimates of serotype-specific transmission model parameters among cattle in the *E. coli* data: the posterior median of the parameter and the 90% credible interval within parentheses. Estimates are multiplied by 100.

Serotype (s)	**Transmission model parameter**			
	ν_s	α_s	β_s	μ_s
A	3.085 (0.835, 5.714)	0.119 (0.055, 0.194)	1.004 (0.444, 1.692)	15.793 (10.234, 22.019)
C	0.710 (0.003, 2.157)	0.054 (0.013, 0.102)	0.622 (0.101, 1.205)	9.102 (3.526, 15.718)
G	1.112 (0.007, 2.898)	0.074 (0.020, 0.135)	1.408 (0.529, 2.509)	17.908 (10.098, 26.893)
M	0.805 (0.003, 2.143)	0.052 (0.011, 0.099)	0.373 (0.009, 0.888)	6.703 (1.004, 17.147)
O	1.785 (0.193, 3.832)	0.149 (0.076, 0.232)	0.722 (0.334, 1.182)	10.533 (6.921, 14.428)
P	1.279 (0.107, 2.854)	0.041 (0.007, 0.081)	1.582 (0.964, 2.247)	11.629 (7.317, 16.067)
T	1.136 (0.013, 2.871)	0.121 (0.053, 0.195)	1.146 (0.554, 1.832)	13.265 (9.128, 17.870)
U	2.136 (0.022, 4.776)	0.187 (0.091, 0.296)	0.725 (0.332, 1.169)	9.452 (6.338, 12.817)

Table 7.6: Estimates of observation model parameters among cattle in the *E. coli* data: the posterior median of the parameter and the 90% credible interval. Estimates are multiplied by 100. See definitions in Section 7.2.

Observation model parameter				
θ_R	θ_F	θ_C	θ_S	$1 - \theta_U$
76.13 (72.80, 79.40)	45.60 (42.30, 48.79)	78.46 (72.44, 84.51)	3.27 (0.37, 5.96)	95.53 (88.53, 99.97)

To explore the effect of our prior specifications, we perform a sensitivity analysis using different hyperparameter values each time. Results are shown in Figure F.4 of the Appendix for the transmission and clearance rate parameters and Figure 7.11 for parameters γ and δ. The posterior distributions of the within-pen and external transmission rates, as well as the clearance rates remained unchanged. However, the use of an uninformative prior, $\text{Exp}\,(0.01)$, leads to an increase in posterior uncertainty for the competition parameter δ. A possible explanation is that our data are only weakly informative due to the relatively small number of type-specific transitions $r \mapsto s$, where $r, s \neq 0$ (posterior median 27, 95% CI [21,35]).

For completeness, we also fitted simpler models to the data, namely a model where a common clearance rate is assumed for the common serotypes and a model for which we set $\delta = 0$. Posterior distributions for the parameters of interest are shown in Figure 7.12. As expected, when a common clearance rate is assumed, the new estimate is approximately equal to the average of the full model estimates for common serotypes but less associated variability. This leads to minor changes in external and within-pen colonisation rates. Moreover, assuming that $\delta = 0$ has only small effects on posterior estimates of the remaining parameters. Possibly, this is due to the very low number of serotype-to-serotype transitions that were estimated in the full model, or it may indicate that parameter δ does not offer substantial improvements in model fit.

7.6 Discussion

In this chapter, we have developed a model for analysing longitudinal household carriage studies with multiple serotypes. Our model improves on existing methodologies by allowing imperfect test sensitivity, that is, that the true carriage states can be falsely recorded as non-carrier or misclassified as another serotype (Cauchemez *et al.*, 2006; Melegaro *et al.*, 2007; Erästö *et al.*, 2012; Numminen *et al.*, 2013). Furthermore, it gains flexibility by allowing non-typed samples to be classified as any of the studied serotypes rather than pulling them into the unidentified group, as it is assumed by the majority of the aforementioned models. Although our method was motivated by a study of repeated observations of *E. coli* colonisation, it can be applied with minor modifications to other infectious diseases such as pneumonococ-

Figure 7.11: Results of our prior sensitivity analysis for the relative colonisation rates.

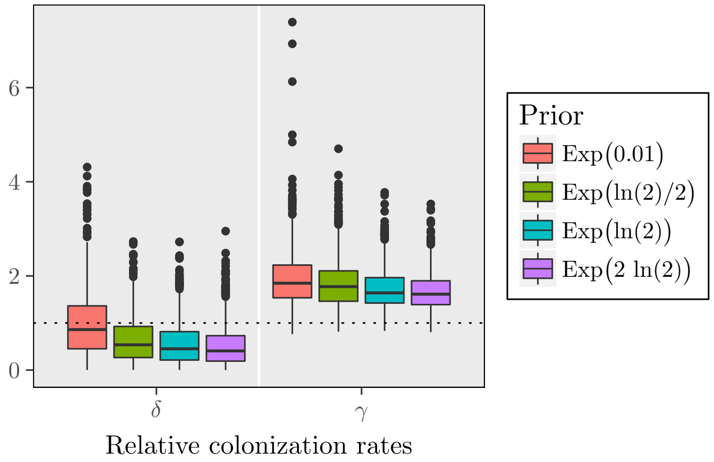
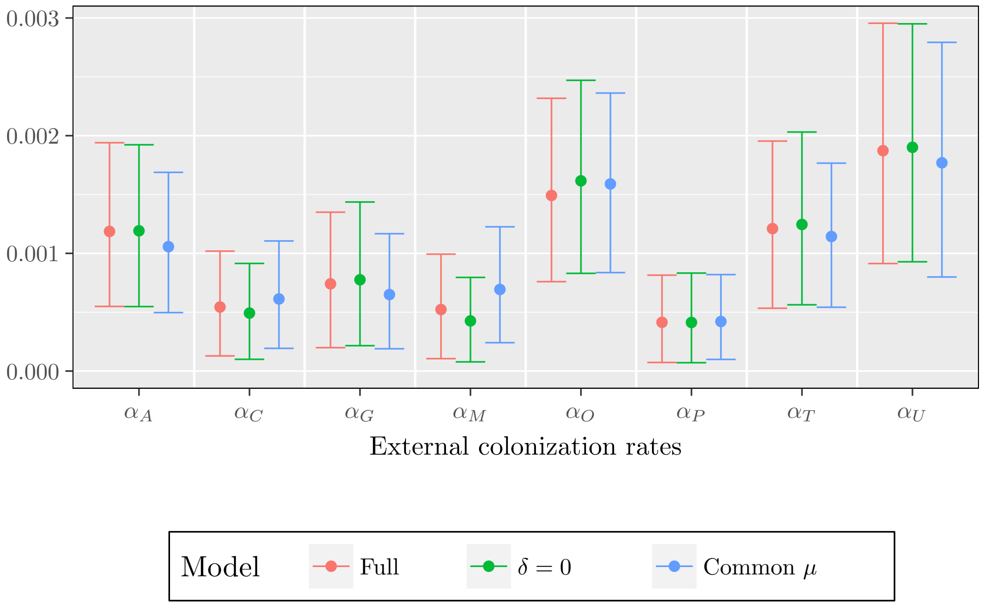
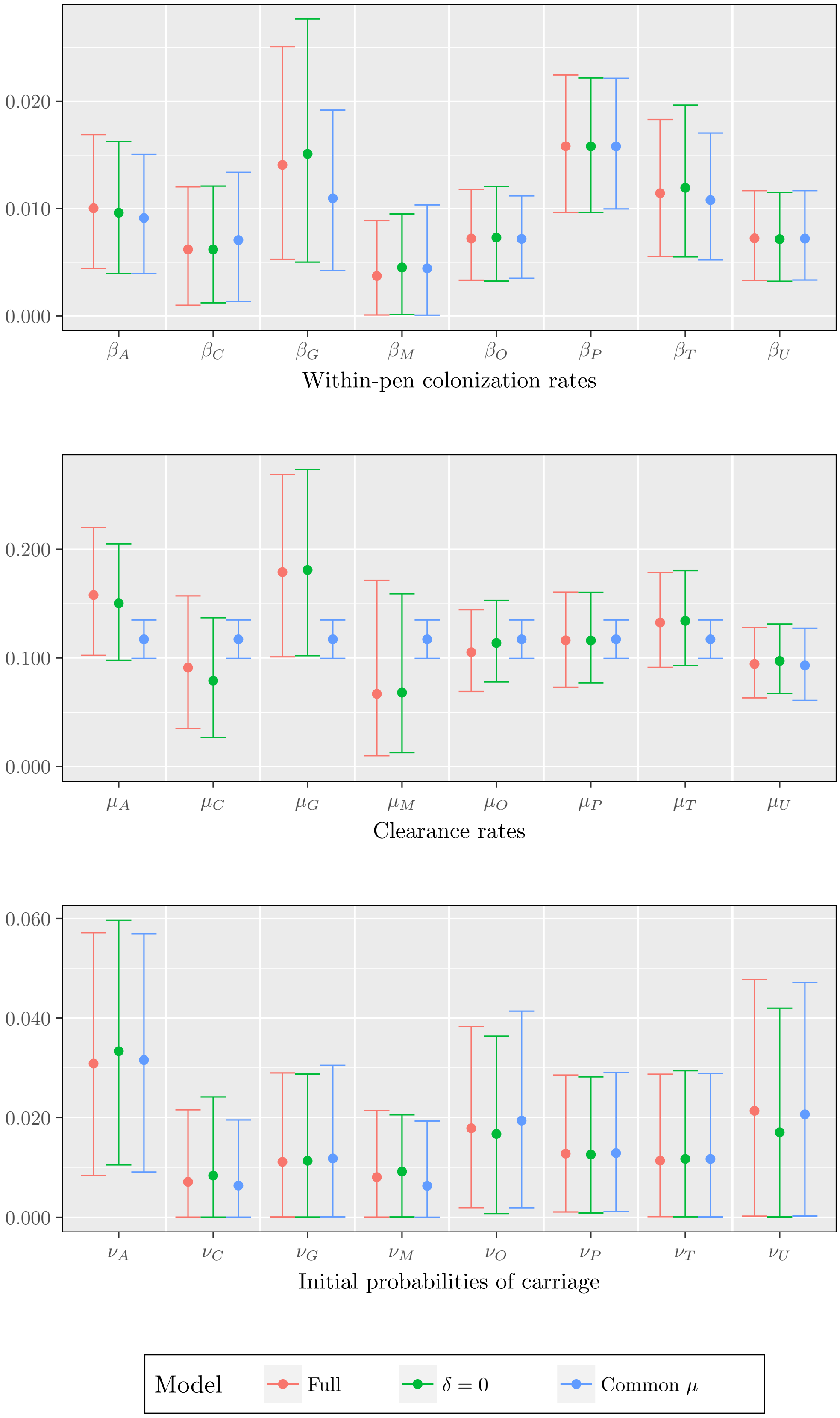
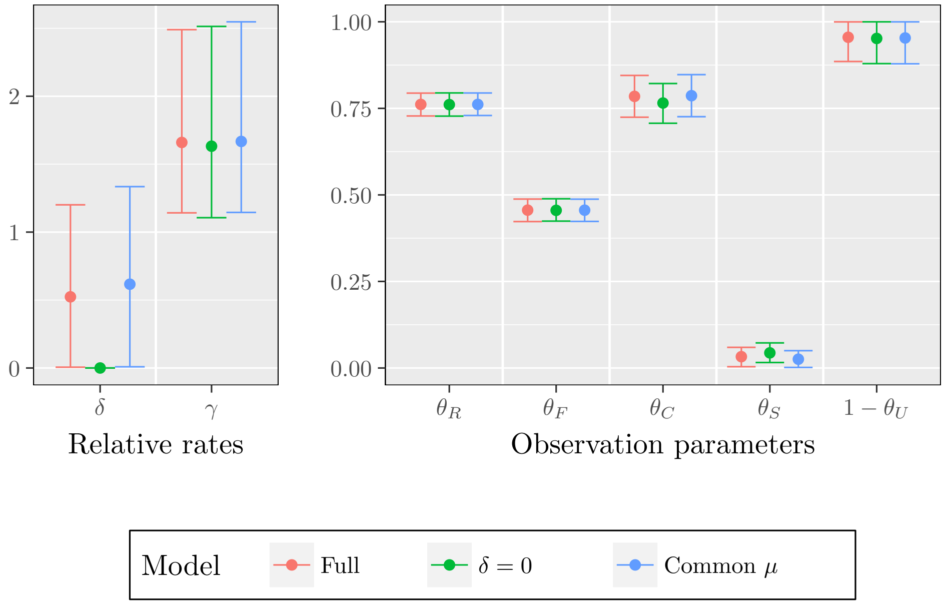


Figure 7.12: Marginal posterior summaries of model parameters among cattle in the *E. coli* data. The median value (dot) and 90% credible interval (error bars) are depicted for each model; red for the full model, green for the model with $\delta = 0$ and blue for the model with a common clearance rate for the common serotypes.




0.020
0.010
0.000
β_A
β_C
β_G
β_M
β_O
β_P
β_T
β_U
Within-pen colonization rates
0.200
0.100
0.000
μ_A
μ_C
μ_G
μ_M
μ_O
μ_P
μ_T
μ_U
Clearance rates
0.060
0.040
0.020
0.000
ν_A
ν_C
ν_G
ν_M
ν_O
ν_P
ν_T
ν_U
Initial probabilities of carriage
Model
Full
$\delta = 0$
Common μ



cus.

The proposed iFFBS algorithm of Section 6.4.4 has allowed us to fit the multi-strain model. An advantage of this approach compared to previous approaches for estimating serotype-specific parameters is that it can be efficiently applied with several serotypes and can reduce correlation between posterior samples by updating the entire true carriage process at each iteration. Moreover, unlike the standard FFBS method, our method does not rely on a small population size or a small total number of different serotypes.

Simulations demonstrate that the algorithm accurately estimates our model parameters and successfully reproduces the incidence of colonisation. A sensitivity analysis was conducted to explore different serotyping strategies. The results indicate that the method performs reasonably well even when only a limited number of serotyped observations are obtained suggesting that simulations can be used in order to design an optimal serotyping scheme in a particular study.

Application of our method to a longitudinal study of *E. coli* in cattle divided in pens has given us valuable insights into the data. Our analysis provides evidence for between-serotype competition since the relative colonisation rate for carriers versus non-carriers was estimated 0.5. Small pens were more susceptible to within-pen colonisation compared to larger pens, as suggested by a relative transmission rate of 1.66.

Differences between serotypes were detected with respect to the external rate of colonisation. In particular, we found that serotype U has the highest rate while serotype P has the lowest. This is expected given that U represents a total of 24 serotypes in the data. For within-pen rate we had significant differences between serotypes M and P, the latter being 4 times as big as the former. Similarities were also observed for serotypes A and T in terms of both external and within-pen colonisation rates. Serotypes M and C share low external rate as well as low within-pen rate. Clearance rates were relatively homogeneous, in the sense that posterior credible intervals for these parameters were overlapping.

A significant merit of our approach that it allows for imperfect diagnostic tests. Using this feature we estimated that in the real data the sensitivity of the faecal test is as low as 46% and the one of the RAMS test is close to 76%. In addition, we concluded that only 80% of the observed common serotypes were correctly identified as the right type. These findings highlight the importance of the imperfect test assumption which has been ignored in several epidemiological studies.

Following previous work on the field, our model treats different serotypes that appear rarely in the observed data as a single group. Even though this is a strong assumption, we believe that this does not impair our inferences since we don't provide any epidemiological interpretation for the transmission parameters of this group. Further, we allow the unidentified group to have its own probability of serotype misclassification.

A potential limitation of our model is that we currently do not allow for co-colonisation, that is, an individual carrying more than one serotypes at a time. Even though this assumption maybe appropriate for *E. coli* O157:H7, it may be unrealistic in other epidemiological studies. Nevertheless, the model can be extended to allow for colonisation by all pairwise combinations of single carriage states and the same algorithm can be used for posterior inferences. Finally, an extension of our model that one may consider is accounting for between pen interactions; this can be achieved by adding an extra between-pen transmission parameter as was done in Section 6.7.2 for a binary state process. However, on account of our findings in Section 5.4.2.1 which suggest no interaction between pens in this particular study, we choose not to consider this extension.

Chapter 8

Conclusions And Extensions

8.1 Summary of the thesis

The main objective of this thesis has been to develop Bayesian techniques for parameter estimation and model comparison in the context of longitudinal household epidemic data. Special emphasis is given to the role of missing data which arise due to partial observation of the true epidemic process and therefore require advanced tools to overcome analytical intractability of the model likelihood. In this section we summarise the main contributions of our work.

In Chapter 3 we present a discrete-time hidden Markov model for infectious disease dynamics within a community of individuals divided into groups. The model allows for two possible routes of transmission, namely within group and transmission from the community. The main advantage of our approach is that it accounts for imperfect tests that are used to detect the disease, by assuming that the data are indirect measurements of a true hidden epidemic process. Parameter estimation is achieved by using MCMC data augmentation methods, where the hidden colonisation states are imputed with the forward filtering backward sampling algorithm. The chapter introduces the modelling and inferential framework upon which the work of subsequent chapters are built.

Despite the methodological advances in parameter estimation, model selection for stochastic epidemic models has been extremely challenging until recently, mainly due to the computationally intensive methods needed to impute the missing times of acquiring and clearing infection. Motivated by this fact, in Chapter 4 we propose a three stage algorithm for efficiently estimating marginal likelihoods in applications with a large amount of missing data. Thus, our approach is well suited for epidemiological problems where each model represents an important hypothesis. The method combines MCMC, importance sampling and filtering to calculate the evidence in favour of each hypothesis considered. An extensive simulation study suggested that the proposed algorithm performs better or equally compared to several alternative approaches, in terms of computational efficiency and accuracy. A significant merit of our approach, that was not exploited in this work, is that it can be easily applied in parallel which can be utilised to speed up implementation, especially in applications involving big data.

The statistical tools developed in Chapter 4 have been applied in Chapter 5 to uncover new insights into the transmission dynamics of *Escherichia coli* O157:H7 in cattle, for which relatively little is known. Based on two longitudinal studies of *E. coli* O157:H7 we were able to demonstrate that contaminated water troughs play an important role in transmission. Furthermore, we have found evidence in favour of a non-Markovian model for the colonisation period of *E. coli* O157:H7, which implies the probability that an individual clears the disease grows as the colonisation period increases. This result may indicate an immune response in the host. Finally, our analyses provide support to the hypothesis that within-pen colonisation rates are higher in pens with smaller area, possibly due to more contacts between individuals that occur in such pens.

In Chapters 3–5 imputation of the missing data has been done with the FFBS algorithm. Even though this method performs very effectively in small scale epidemics, it can be computationally prohibitive as the total number of subjects per household increases or in applications with multi-type hidden states. Chapter 6 introduces a novel extension of the FFBS algorithm, called iFFBS, that is suitable for inference on large scale epidemic data. The iFFBS algorithm updates the hidden infection states individually per subject conditional on the rest, as opposed to the standard FFBS algorithm where sampling is done for all individuals jointly. The computation time for the iFFBS is linear in the population size rather than the exponential time needed for the FFBS. The approach relies on a Markovian colonisation period, but it is shown that this assumption can be relaxed to extend the iFFBS to non-Markovian models as well. Moreover, our method has been successfully applied in a stochastic epidemic model with three possible routes of transmission, where each individual belongs to a primary group (e.g. household), a secondary group (e.g. neighbourhood) and the community. In all of these applications, we found that iFFBS outperforms existing approaches in terms of computational efficiency and scales well with dimensionality.

In Chapter 7 we developed an individual-based multi-state model for the transmission dynamics of *E. coli* O157:H7 serotypes. To our knowledge, this is the first study that accounts for both imperfect diagnostic tests and serotype misclassification. Bayesian inference for this model is performed using the iFFBS algorithm presented in Chapter 6, providing accurate estimates of the model parameters and the true carriage process in simulated data. Application on a real dataset provided valuable insights regarding the dynamics of different strains of *E. coli* O157:H7. In particular, our findings suggest between serotype competition and various heterogeneities in transmission rates between different serotypes. To conclude, we highlight that even though the work presented in this thesis has mainly considered *E. coli* O157:H7, the statistical methodology is flexible and can be applied to a wide range of other infectious diseases.

8.2 Extensions

The importance sampling method for model comparison proposed in Chapter 4 uses the forward filtering backward sampling algorithm to construct the proposal density of the hidden colonisation process. The computational cost of this step grows exponentially in the number of individuals per household, and therefore can only be applied in small population problems. In our future work, we aim to improve the computational efficiency of our approach in order to be applicable to large populations. One possible solution to this problem is using the iFFBS method developed in Chapter 6 instead of the FFBS algorithm.

In Chapter 5 we have proposed an MCMC algorithm for inference in discrete time Susceptible-Infected-Susceptible models with a Negative Binomially distributed colonisation period. Our approach is based on a Metropolis Hasting update for the hidden epidemic process, with proposal density being the full conditional under the Geometric model. Therefore, the accuracy of the method decreases as the dispersion parameter of the Negative Binomial distribution moves away from 1. In our future work we will investigate alternative updates for this model. One way to address the problem is to use ideas from Tokdar *et al.* (2010), who use an adapted version of the FFBS algorithm after expanding the original state space so that the hidden semi-Markov model takes a Markovian form; this is achieved by introducing a duration variable. The computational complexity of this algorithm increases as the study period grows, since at every step we need to sum over the possible duration intervals, as well as the number of hidden states. Alternatively, we can follow Langrock *et al.* (2012) who extended the hidden state space into a set of N^* possible states, defining suitable transition probabilities in order to construct a hidden Markov model that approximates the hidden semi-Markov model. As before, this approach is more computationally demanding compared to our approach.

There are several ways in which the iFFBS method can be further speeded up. In its current form, the algorithm iteratively simulates the unobserved carriage process of all individuals in the study. An interesting question is whether it is necessary to update all individuals at every iteration of the algorithm or if it is more computationally efficient to choose a subset of the individuals. One could consider various strategies as to how these subsets are chosen. For instance, a proportion of individuals can be selected uniformly at random. An alternative approach would be to update the hidden states of each individual c with probability $p^{[c]}$, where $p^{[c]}$ is the proportion of iterations of a pilot iFFBS run in which the colonisation states changed compared to the previous iteration. Intuitively, we expect that mixing of the Markov chains in the latter to be better compared to the former, even though the computational cost is higher. In the future, we will investigate which one of these approaches is optimal.

The multi-serotype model presented in Chapter 7 could be extended in many ways. In particular, we have assumed a Geometric distribution of the colonisation period and this assumption could be relaxed by using the methodology of Chapter 6 for hidden semi-Markov models. Furthermore, the model can be modified to account for neighbouring pen interactions by adding an additional parameter for between-pen transmission and inference can be carried out as shown in Section 6.7.2. Finally, we will consider relaxing our assumption of single serotype carriage, thus allowing for an individual to be colonised with multiple serotypes at a time termed co-infection. The issue of co-infection has recently been understood to be of great importance in epidemiology, due to potential for the exchange of genetic material within co-colonised individuals leading to antigenic shifts in the pathogen. We anticipate that statistical tools such as our proposed iFFBS algorithm will shortly be in great demand as datasets emerge that track co-infection. This can be easily done under the current framework by considering all pairwise serotype combinations as additional possible states. However we note that in all of the aforementioned extensions, the increase in parameters may lead to identifiability issues which could be tackled by assuming some common serotype parameters.

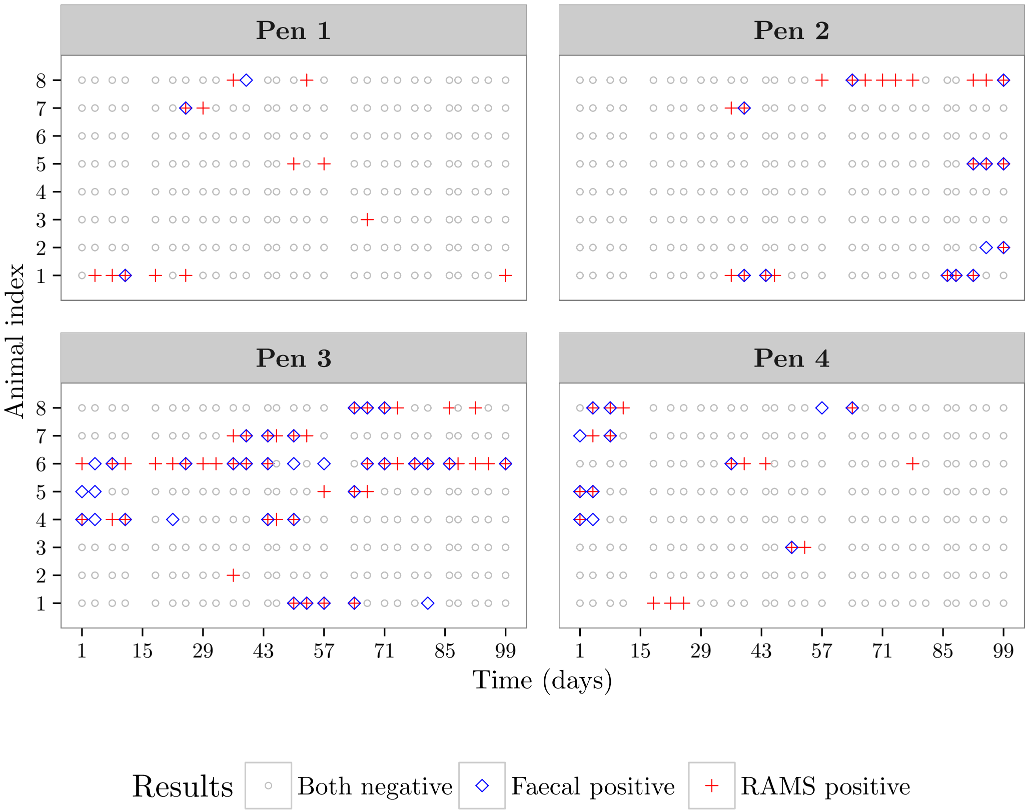
APPENDIX A

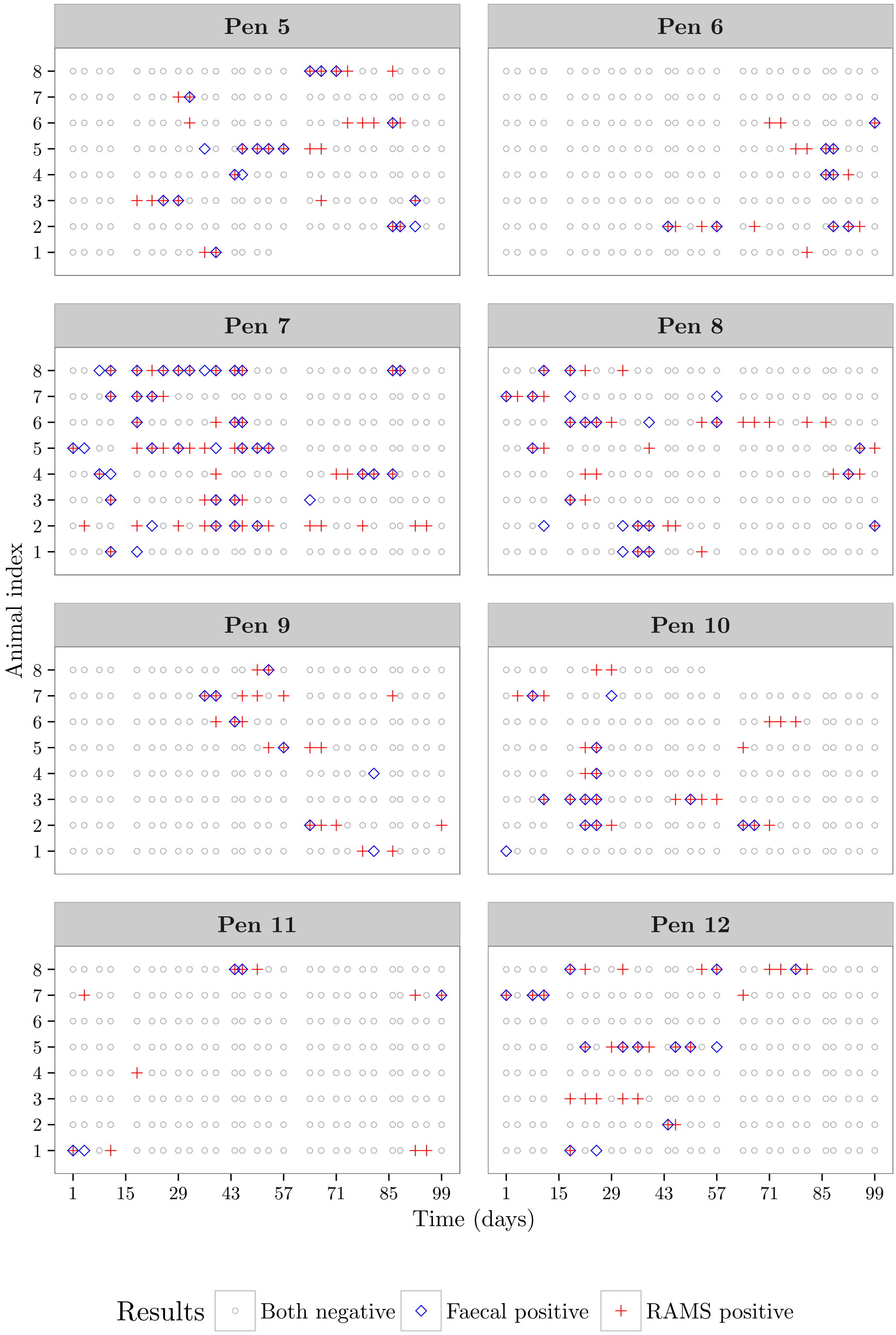
MOTIVATING DATASETS SUPPLEMENTARY MATERIAL

A.1 Patterns of colonisation

Figure A.1 and Figure A.2 show the plots of observed data collected from *E. coli* O157:H7 dataset 1 and *E. coli* O157:H7 dataset 2 respectively.

FIGURE A.1: RAMS and faecal samples collected in the 20 pens participating in the first *E. coli* O157:H7 longitudinal study. Test results stated as positive; "+" when the RAMS test is positive and "◇" when the faecal test is positive. Grey circle ("○") indicates that the sample is taken but no *E. coli* O157:H7 is detected by both tests. Missing data between the samplings days is indicated by the white space between the points.




Pen 5
Pen 6
Pen 7
Pen 8
Pen 9
Pen 10
Pen 11
Pen 12
Animal index
1
2
3
4
5
6
7
8
Time (days)
1
15
29
43
57
71
85
99
Results
○ Both negative
◇ Faecal positive
+ RAMS positive

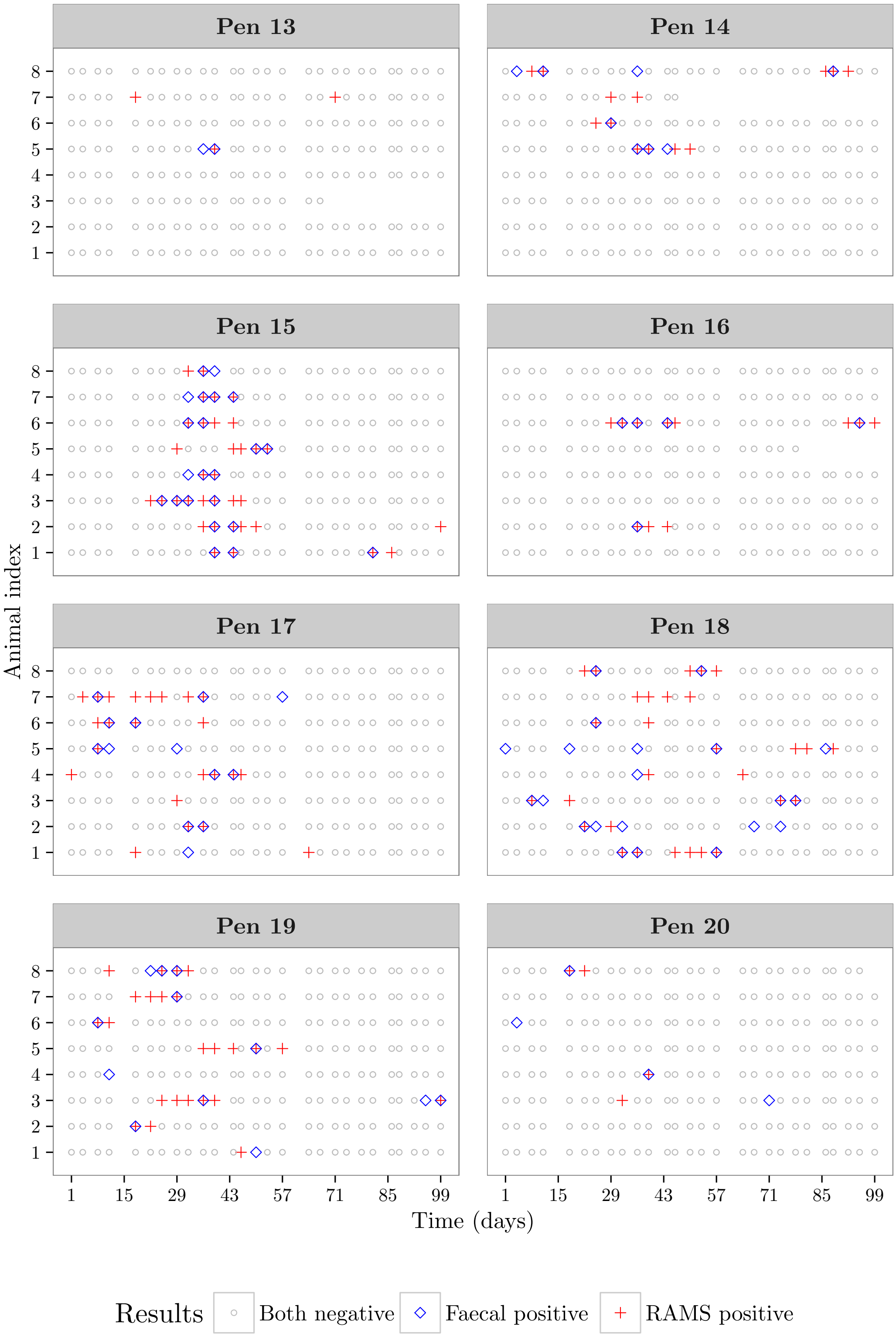
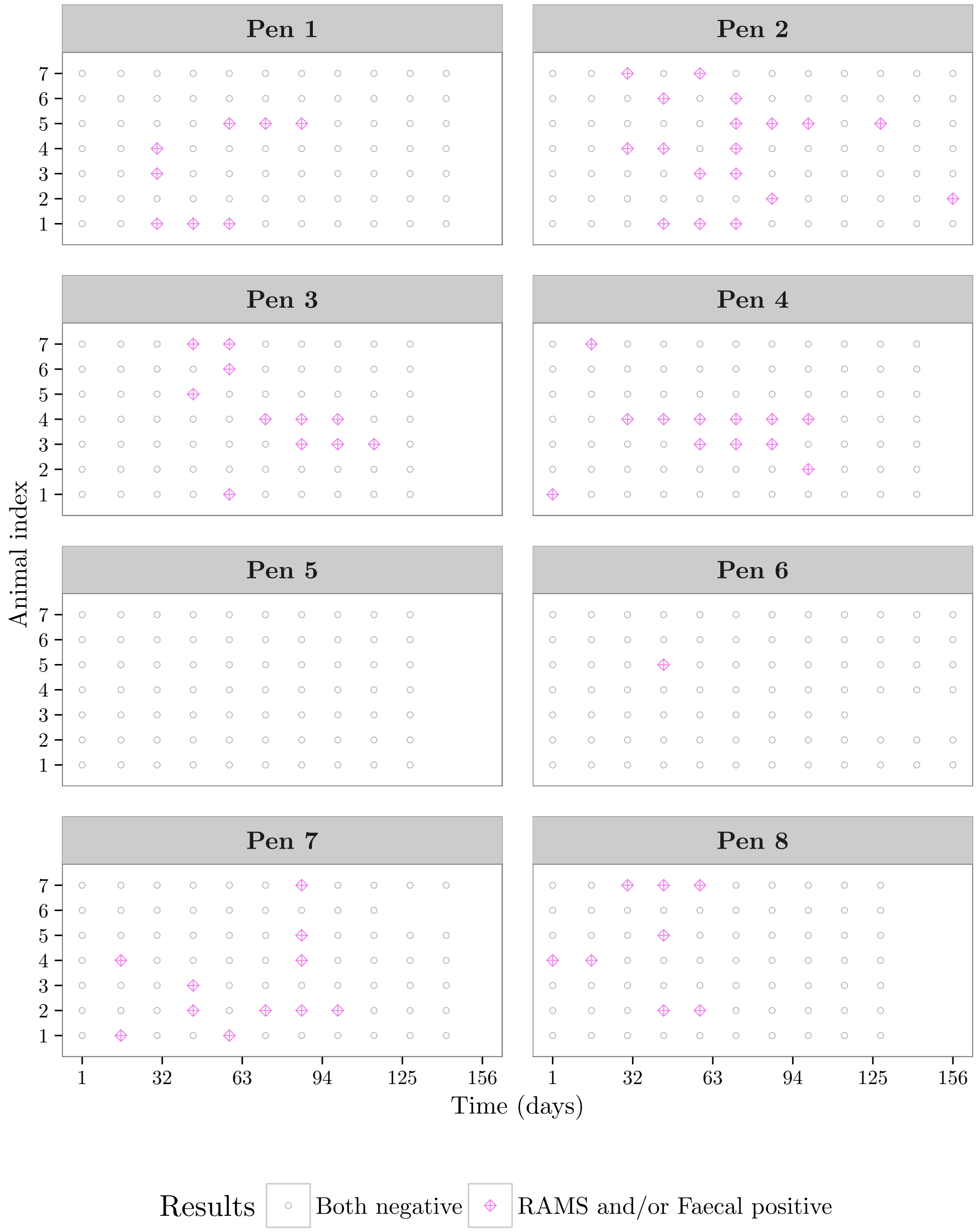
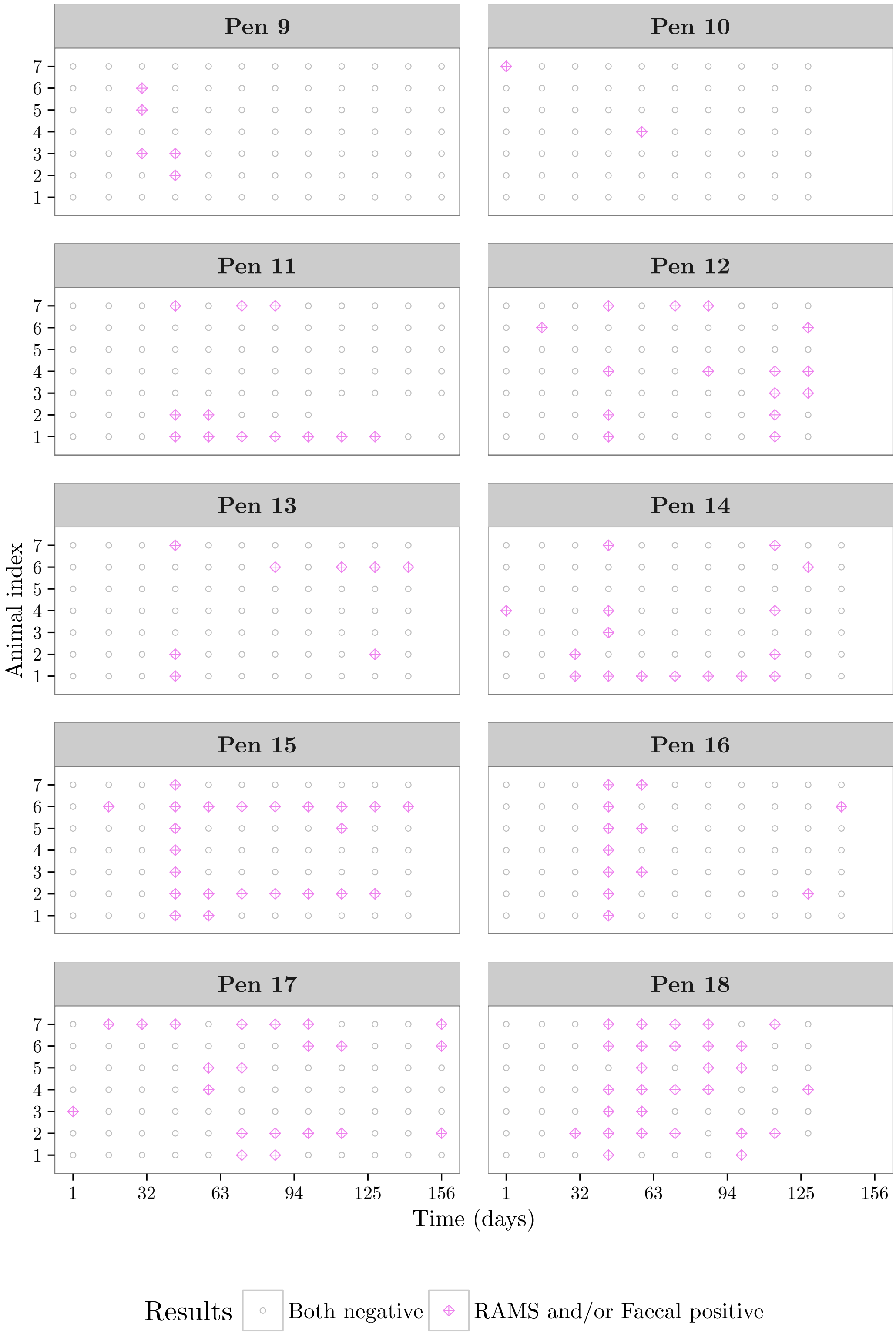
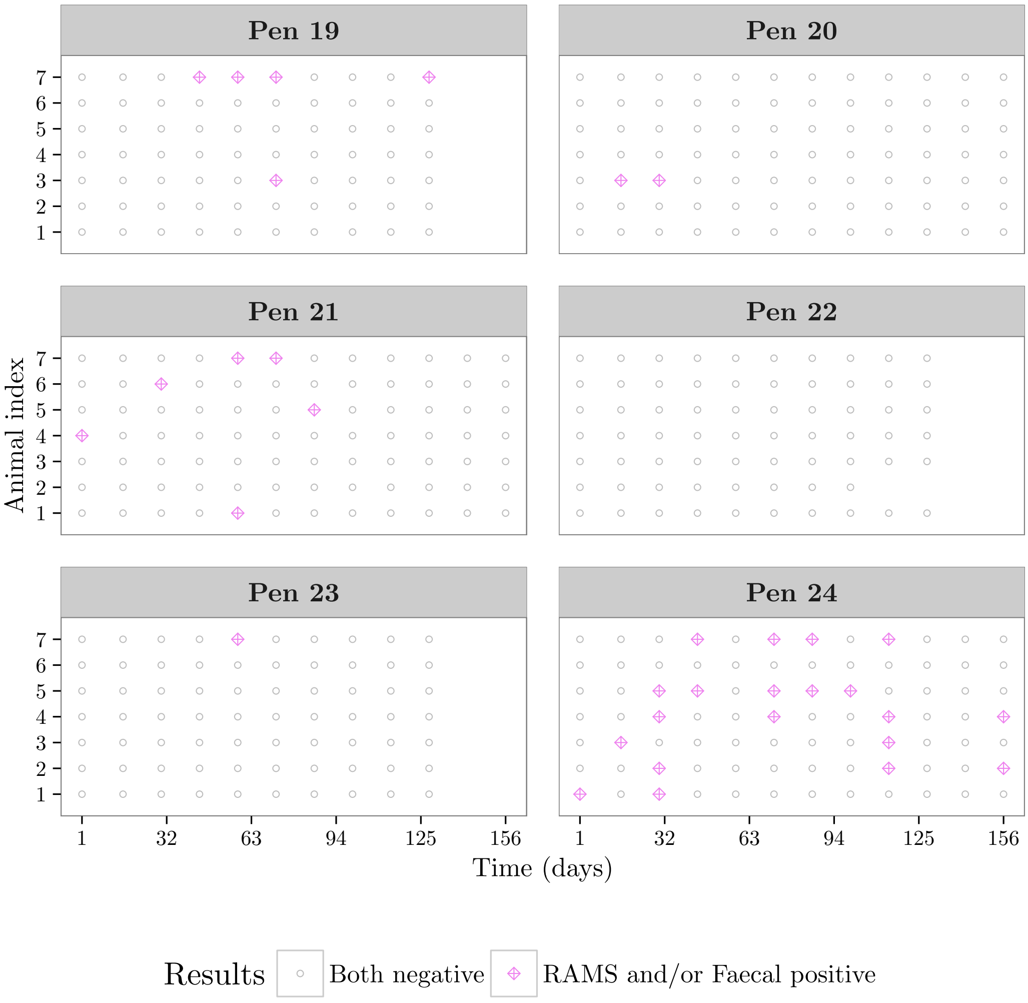

Pen 13
Pen 14
Pen 15
Pen 16
Pen 17
Pen 18
Pen 19
Pen 20
Animal index
1
2
3
4
5
6
7
8
Time (days)
1
15
29
43
57
71
85
99
Results
○ Both negative
◇ Faecal positive
+ RAMS positive

Figure A.2: RAMS and faecal samples collected in the 24 pens participating in the second *E. coli* longitudinal study. Test results stated as positive; "⊕" when the RAMS and/or faecal test is positive and "○" indicates that the sample is taken but no *E. coli* O157:H7 is detected by both tests. Missing data between the samplings days is indicated by the white space between the points.




Pen 9
Pen 10
Pen 11
Pen 12
Pen 13
Pen 14
Pen 15
Pen 16
Pen 17
Pen 18
Animal index
1
2
3
4
5
6
7
Time (days)
1
32
63
94
125
156
Results
○ Both negative
⊕ RAMS and/or Faecal positive


Pen 19
Pen 20
Pen 21
Pen 22
Pen 23
Pen 24
Animal index
1
2
3
4
5
6
7
Time (days)
1
32
63
94
125
156
Results
Both negative
RAMS and/or Faecal positive

Appendix B

Hidden Markov Model For Household Epidemic Data Supplementary Material

B.1 MCMC details

In this section we provide the details of the MCMC Algorithm 3 presented in Section 3.4, used to generate samples from the posterior of the basic HMM of Chapter 3. Simulation of the hidden states $\mathbf{X}$ has been already discussed in the main body. We now give details for the remaining model parameters.

B.1.1 Updating the transmission parameters

The full conditional distribution of the initial colonisation parameter ν, is:

$$\begin{aligned}\pi(\nu \mid \mathbf{Y}, \mathbf{X}, \tilde{\alpha}, \tilde{\beta}, \tilde{m}, \theta_R, \theta_F) &\propto \nu^{b_\nu - 1}(1-\nu)^{c_\nu - 1} \prod_{p=1}^{P}\prod_{c=1}^{n_c^p}\left[\nu^{x_1^{[c,\,p]}}(1-\nu)^{1-x_1^{[c,\,p]}}\right] \\ &= \nu^{\sum_{p=1}^{P} \sum_{c=1}^{n_c^p} x_1^{[c,\,p]} + b_\nu - 1}(1-\nu)^{\sum_{p=1}^{P}\sum_{c=1}^{n_c^p}\left(1-x_1^{[c,\,p]}\right)+c_\nu-1}.\end{aligned}$$

Hence we draw $\nu \mid \cdot \sim \text{Beta}\left(\sum_{p=1}^{P}\sum_{c=1}^{n_c^p} x_1^{[c,\,p]} + b_\nu, \sum_{p=1}^{P}\sum_{c=1}^{n_c^p}\left(1 - x_1^{[c,\,p]}\right) + c_\nu\right)$. For $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{m}$ the joint full conditional is given up to a multiplicative constant as:

$$\begin{aligned}&\pi(\tilde{\alpha}, \tilde{\beta}, \tilde{m} \mid \mathbf{Y}, \mathbf{X}, \nu, \theta_R, \theta_F) \\ &\quad\propto \prod_{p=1}^{P}\prod_{t=2}^{T^p}\left[\left(\exp\left\{-e^{\tilde{\alpha}} - e^{\tilde{\beta}}\sum_{c=1}^{n_c^p} x_{t-1}^{[c,\,p]}\right\}\right)^{N_{00}^p(t)}\left(\frac{\tilde{m}}{\tilde{m}+1}\right)^{N_{11}^p(t)}\left(\frac{1}{\tilde{m}+1}\right)^{N_{10}^p(t)}\right. \\ &\qquad\left.\times\left(1-\exp\left\{-e^{\tilde{\alpha}} - e^{\tilde{\beta}}\sum_{c=1}^{n_c^p} x_{t-1}^{[c,\,p]}\right\}\right)^{N_{01}^p(t)}\right] \times \pi_\alpha\big(e^{\tilde{\alpha}}\big) \times e^{\tilde{\alpha}} \times \pi_\beta\big(e^{\tilde{\beta}}\big) \times e^{\tilde{\beta}} \times \pi_{\tilde{m}}(\tilde{m}),\end{aligned}$$

where $N^p_{kj}(t)$ the number of individuals in pen p who were in state k at time $t-1$ and in state j at time t, for $k, j \in \{0, 1\}$.

This distribution cannot be solved analytically and therefore we use HMC to update these parameters. We have that the partial derivatives are given by:

$$\frac{\partial \log \pi(\tilde{\alpha}, \tilde{\beta}, \tilde{m} \mid \cdot)}{\partial \tilde{\alpha}} = \sum_{p=1}^{P} \sum_{t=2}^{T^p} \left[N^p_{01}(t) \times e^{\tilde{\alpha}} \times \frac{\exp\left\{ - e^{\tilde{\alpha}} - e^{\tilde{\beta}} \sum_{c=1}^{n^p_c} x^{[c,p]}_{t-1} \right\}}{1 - \exp\left\{ - e^{\tilde{\alpha}} - e^{\tilde{\beta}} \sum_{c=1}^{n^p_c} x^{[c,p]}_{t-1} \right\}} - N^p_{00}(t) \times e^{\tilde{\alpha}} \right] + \frac{\partial \log \pi_\alpha(e^{\tilde{\alpha}})}{\partial \tilde{\alpha}} + 1,$$

$$\frac{\partial \log \pi(\tilde{\alpha}, \tilde{\beta}, \tilde{m} \mid \cdot)}{\partial \tilde{\beta}} = \sum_{p=1}^{P} \sum_{t=2}^{T^p} \left[N^p_{01}(t) \times e^{\tilde{\beta}} \times \sum_{c=1}^{n^p_c} x^{[c,p]}_{t-1} \times \frac{\exp\left\{ - e^{\tilde{\alpha}} - e^{\tilde{\beta}} \sum_{c=1}^{n^p_c} x^{[c,p]}_{t-1} \right\}}{1 - \exp\left\{ - e^{\tilde{\alpha}} - e^{\tilde{\beta}} \sum_{c=1}^{n^p_c} x^{[c,p]}_{t-1} \right\}} - N^p_{00}(t) \times e^{\tilde{\beta}} \times \sum_{c=1}^{n^p_c} x^{[c,p]}_{t-1} \right] + \frac{\partial \log \pi_\beta(e^{\tilde{\beta}})}{\partial \tilde{\beta}} + 1,$$

$$\frac{\partial \log \pi(\tilde{\alpha}, \tilde{\beta}, \tilde{m} \mid \cdot)}{\partial \tilde{m}} = \sum_{p=1}^{P} \sum_{t=2}^{T^p} \left[\frac{N^p_{11}(t)}{\tilde{m}} + \frac{N^p_{11}(t) + N^p_{10}(t)}{\tilde{m} + 1} \right] + \frac{\partial \log \pi_{\tilde{m}}(\tilde{m})}{\partial \tilde{m}}.$$

We use a fixed number of leapfrog steps $L = 30$ and adopt the stepsize ϵ during burn-in to obtain an acceptance rate of roughly 65% as suggested by Neal (2011).

B.1.2 Updating the observation parameters

The full conditional of θ_R is:

$$\pi(\theta_R \mid \mathbf{Y}, \mathbf{X}, \tilde{\alpha}, \tilde{\beta}, \tilde{m}, \nu, \theta_F) \propto \theta_R^{b_{\theta_R}-1} (1 - \theta_R)^{c_{\theta_R}-1} \prod_{p=1}^{P} \prod_{c=1}^{n^p_c} \prod_{\substack{x^{[c,p]}_t = 1 \\ t \in O^{c,p}}} \left[(\theta_R)^{\mathrm{r}^{[c,p]}_t} (1 - \theta_R)^{1 - \mathrm{r}^{[c,p]}_t} \right],$$

and for θ_F it is:

$$\pi(\theta_F \mid \mathbf{Y}, \mathbf{X}, \tilde{\alpha}, \tilde{\beta}, \tilde{m}, \nu, \theta_R)$$
$$\propto \theta_F^{b_{\theta_F}-1}(1-\theta_F)^{c_{\theta_F}-1} \prod_{p=1}^{P}\prod_{c=1}^{n_c^p} \prod_{\substack{x_t^{[c,\,p]}=1 \\ t\in O^{c,\,p}}} \left[\left(\theta_F\right)^{\mathrm{f}_t^{[c,\,p]}}\left(1-\theta_F\right)^{1-\mathrm{f}_t^{[c,\,p]}}\right].$$

Hence we draw θ_R and θ_F:

$$\theta_R \mid \cdot \quad \sim \quad \text{Beta}\left(\sum_{p=1}^{P}\sum_{c=1}^{n_c^p}\sum_{\substack{x_t^{[c,\,p]}=1 \\ t\in O^{c,\,p}}} \mathrm{r}_t^{[c,\,p]} + b_{\theta_R}, \sum_{p=1}^{P}\sum_{c=1}^{n_c^p}\sum_{\substack{x_t^{[c,\,p]}=1 \\ t\in O^{c,\,p}}} \left(1-\mathrm{r}_t^{[c,\,p]}\right) + c_{\theta_R}\right),$$

$$\theta_F \mid \cdot \quad \sim \quad \text{Beta}\left(\sum_{p=1}^{P}\sum_{c=1}^{n_c^p}\sum_{\substack{x_t^{[c,\,p]}=1 \\ t\in O^{c,\,p}}} \mathrm{f}_t^{[c,\,p]} + b_{\theta_F}, \sum_{p=1}^{P}\sum_{c=1}^{n_c^p}\sum_{\substack{x_t^{[c,\,p]}=1 \\ t\in O^{c,\,p}}} \left(1-\mathrm{f}_t^{[c,\,p]}\right) + c_{\theta_F}\right).$$

B.2 Handling individual dropouts

In both of our datasets we observe some occasions where exactly one individual c_d^p leaves the study before completion of its pen observation period T^p. Let $T^{[c_d^p,\,p]}$ be the day of the last observation that we obtain for this individual. Note that in this case, $\mathbf{X}^p$ does not include the carriage states of c_d^p after time $T^{[c_d^p,\,p]}$. After the dropout, the individual does not play any role in the epidemic process and hence the probability of $\mathbf{X}^p \mid \tilde{\alpha}, \tilde{\beta}, \tilde{m}, \nu$ is written as:

$$\begin{aligned}\mathbb{P}\left(\mathbf{X}^p = \mathbf{x}^p \mid \tilde{\alpha}, \tilde{\beta}, \tilde{m}, \nu\right) = & \prod_{c=1}^{n_c^p}\left(\mathbb{P}\left(X_1^{[c,\,p]} = x_1^{[c,\,p]} \mid \nu\right)\prod_{t=2}^{T^{[c_d^p,\,p]}} P^{[c,\,p]}_{x_{t-1}^{[c,\,p]},\, x_t^{[c,\,p]},\, t}\right) \\ & \times \prod_{\substack{c=1 \\ c\neq c_d^p}}^{n_c^p}\left(P^{[c,\,p]}_{x^{[c,\,p]}_{T^{[c_d^p,\,p]}},\, x^{[c,\,p]}_{T^{[c_d^p,\,p]}+1},\, T^{[c_d^p,\,p]}+1}\right) \\ & \times \prod_{\substack{c=1 \\ c\neq c_d^p}}^{n_c^p}\prod_{t=T^{[c_d^p,\,p]}+2}^{T^p} \tilde{P}^{[c,\,p]}_{x_{t-1}^{[c,\,p]},\, x_t^{[c,\,p]},\, t},\end{aligned}$$

where $\tilde{P}$ is defined as the transition probability excluding the contribution of individual c_d^p. When population sizes vary we need to specify how transmission probabilities change. In this cases, we assume density dependent transmission. This change is accounted in the update of the model parameters. For the hidden carriage process, we modify the FFBS algorithm as follows:

1. Initialise the forward recursion at $t = 1$:

$$\mathbb{P}\left(\mathbf{X}_1^{[1:n_c^p,\,p]} = \mathbf{k} \mid \mathbf{Y}_1^{[1:n_c^p,\,p]}, \tilde{\boldsymbol{\theta}}\right) = \frac{\mathbb{P}\left(\mathbf{X}_1^{[1:n_c^p,\,p]} = \mathbf{k} \mid \nu\right) f_{\mathbf{k}}\left(\mathbf{y}_1^{[1:n_c^p,\,p]} \mid \boldsymbol{\vartheta}\right)}{\sum\limits_{\boldsymbol{\omega}\,\in\,\mathcal{X}_s^{n_c^p}} \mathbb{P}\left(\mathbf{X}_1^{[1:n_c^p,\,p]} = \boldsymbol{\omega} \mid \nu\right) f_{\boldsymbol{\omega}}\left(\mathbf{y}_1^{[1:n_c^p,\,p]} \mid \boldsymbol{\vartheta}\right)},$$

   for $\mathbf{k} \in \mathcal{X}_s^{n_c^p}$.

2. For $t = 2, 3, \ldots, T^{[c_d^p,\,p]}$ in a forward recursion, for $\mathbf{k} \in \mathcal{X}_s^{n_c^p}$:

   (a) Compute the one-step ahead predictive probabilities,

$$\begin{aligned}&\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,\,p]} = \mathbf{k} \mid \mathbf{Y}_{1:t-1}^{[1:n_c^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)\\ &\quad= \sum_{\boldsymbol{\omega}\,\in\,\mathcal{X}_s^{n_c^p}} \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,\,p]} = \mathbf{k} \mid \mathbf{X}_{t-1}^{[1:n_c^p,\,p]} = \boldsymbol{\omega}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_{t-1}^{[1:n_c^p,\,p]} = \boldsymbol{\omega} \mid \mathbf{Y}_{1:t-1}^{[1:n_c^p,\,p]}, \tilde{\boldsymbol{\theta}}\right).\end{aligned}$$

   (b) Compute the filtered probabilities,

$$\begin{aligned}&\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,\,p]} = \mathbf{k} \mid \mathbf{Y}_{1:t}^{[1:n_c^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)\\ &\qquad= \frac{f_{\mathbf{k}}\left(\mathbf{y}_t^{[1:n_c^p,\,p]} \mid \boldsymbol{\vartheta}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,\,p]} = \mathbf{k} \mid \mathbf{Y}_{1:t-1}^{[1:n_c^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)}{\sum\limits_{\boldsymbol{\omega}\,\in\,\mathcal{X}_s^{n_c^p}} f_{\boldsymbol{\omega}}\left(\mathbf{y}_t^{[1:n_c^p,\,p]} \mid \boldsymbol{\vartheta}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p,\,p]} = \boldsymbol{\omega} \mid \mathbf{Y}_{1:t-1}^{[1:n_c^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)}.\end{aligned}$$

3. For $t = T^{[c_d^p,\,p]} + 1$ and $\tilde{\mathbf{k}} \in \mathcal{X}_s^{n_c^p - 1} = \{0, 1\}^{n_c^p - 1}$:

   (a) Compute the one-step ahead predictive probabilities,

$$\begin{aligned}&\mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{Y}_{1:t-1}^{[1:n_c^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)\\ &\quad= \sum_{\boldsymbol{\omega}\,\in\,\mathcal{X}_s^{n_c^p}} \mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{X}_{t-1}^{[1:n_c^p,\,p]} = \boldsymbol{\omega}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_{t-1}^{[1:n_c^p,\,p]} = \boldsymbol{\omega} \mid \mathbf{Y}_{1:t-1}^{[1:n_c^p,\,p]}, \tilde{\boldsymbol{\theta}}\right),\end{aligned}$$

   where $\mathbf{X}_t^{[-c_d^p,\,p]}$ denotes the vector $\mathbf{X}_t^{[1:n_c^p,\,p]}$ excluding $X_t^{[c_d^p,\,p]}$, and

$$\mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{X}_{t-1}^{[1:n_c^p,\,p]} = \boldsymbol{\omega}, \tilde{\boldsymbol{\phi}}\right) = \prod_{\substack{c=1\\ c\neq c_d^p}}^{n_c^p} P^{[c,\,p]}_{\omega^{[c]},\,\tilde{k}^{[c]},\,t}.$$

(b) Compute the filtered probabilities,

$$\begin{aligned}&\mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{Y}_{1:t}^{[-c_d^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)\\ &\qquad = \frac{\tilde{f}_{\tilde{\mathbf{k}}}\left(\mathbf{y}_t^{[-c_d^p,\,p]} \mid \boldsymbol{\vartheta}\right) \mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{Y}_{1:t-1}^{[-c_d^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)}{\sum\limits_{\tilde{\boldsymbol{\omega}} \,\in\, \mathcal{X}_s^{n_c^p-1}} \tilde{f}_{\tilde{\boldsymbol{\omega}}}\left(\mathbf{y}_t^{[-c_d^p,\,p]} \mid \boldsymbol{\vartheta}\right) \mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\boldsymbol{\omega}} \mid \mathbf{Y}_{1:t-1}^{[-c_d^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)},\end{aligned}$$

where,

$$\tilde{f}_{\tilde{\mathbf{k}}}\left(\mathbf{y}_t^{[-c_d^p,\,p]} \mid \boldsymbol{\vartheta}\right) := \pi\left(\mathbf{Y}_t^{[-c_d^p,\,p]} \mid \mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}}, \boldsymbol{\vartheta}\right) = \prod_{\substack{c=1\\ c\neq c_d^p}}^{n_c^p} f_{\tilde{k}^{[c]}}\left(y_t^{[c,\,p]} \mid \boldsymbol{\vartheta}\right).$$

4. For $t = T^{[c_d^p,\,p]}+2, T^{[c_d^p,\,p]}+3, \ldots, T^p$ in a forward recursion, for $\tilde{\mathbf{k}} \in \mathcal{X}_s^{n_c^p-1}$:

(a) Compute the one-step ahead predictive probabilities,

$$\begin{aligned}&\mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{Y}_{1:t-1}^{[-c_d^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)\\ &\quad = \sum_{\tilde{\boldsymbol{\omega}} \,\in\, \mathcal{X}_s^{n_c^p-1}} \mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{X}_{t-1}^{[-c_d^p,\,p]} = \tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_{t-1}^{[-c_d^p,\,p]} = \tilde{\boldsymbol{\omega}} \mid \mathbf{Y}_{1:t-1}^{[-c_d^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)\end{aligned}$$

where,

$$\begin{aligned}&\mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{X}_{t-1}^{[-c_d^p,\,p]} = \tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\phi}}\right)\\ &\quad = \prod_{\substack{c=1\\ c\neq c_d^p}}^{n_c^p} \mathbb{P}\left(X_t^{[c,\,p]} = \tilde{k}^{[c]} \mid X_{t-1}^{[c,\,p]} = \tilde{i}^{[c]}, \mathbf{X}_{t-1}^{[-(c,\,c_d^p),\,p]} = \tilde{\omega}^{[-c]}, \tilde{\boldsymbol{\phi}}\right) = \prod_{\substack{c=1\\ c\neq c_d^p}}^{n_c^p} \tilde{P}^{[c,\,p]}_{\tilde{\omega}^{[c]},\,\tilde{k}^{[c]},\,t}.\end{aligned}$$

(b) Compute the filtered probabilities,

$$\begin{aligned}&\mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{Y}_{1:t}^{[-c_d^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)\\ &\qquad = \frac{\tilde{f}_{\tilde{\mathbf{k}}}\left(\mathbf{y}_t^{[-c_d^p,\,p]} \mid \boldsymbol{\vartheta}\right) \mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{Y}_{1:t-1}^{[-c_d^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)}{\sum\limits_{\tilde{\boldsymbol{\omega}} \,\in\, \mathcal{X}_s^{n_c^p-1}} \tilde{f}_{\tilde{\boldsymbol{\omega}}}\left(\mathbf{y}_t^{[-c_d^p,\,p]} \mid \boldsymbol{\vartheta}\right) \mathbb{P}\left(\mathbf{X}_t^{[-c_d^p,\,p]} = \tilde{\boldsymbol{\omega}} \mid \mathbf{Y}_{1:t-1}^{[-c_d^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)}.\end{aligned}$$

5. Simulate a value for $\mathbf{X}_{T^p}^{[-c_d^p,\,p]}$ according to the filtered state probabilities $\mathbb{P}\left(\mathbf{X}_{T^p}^{[-c_d^p,\,p]} = \tilde{\mathbf{k}} \mid \mathbf{Y}_{1:T^p}^{[-c_d^p,\,p]}, \tilde{\boldsymbol{\theta}}\right)$, $\tilde{\mathbf{k}} \in \mathcal{X}_s^{n_c^p-1}$.

6. For $t = T^p - 1, T^p - 2, T^{[c_d^p, p]} + 1$ compute the conditional probabilities $\mathbb{P}\left(\mathbf{X}_t^{[-c_d^p, p]} = \tilde{\mathbf{k}} \mid \mathbf{X}_{t+1}^{[-c_d^p, p]} = \mathbf{x}_{t+1}^{[-c_d^p, p]}, \mathbf{Y}_{1:t}^{[-c_d^p, p]}, \tilde{\boldsymbol{\theta}}\right)$, $\tilde{\mathbf{k}} \in \mathcal{X}_s^{n_c^p - 1}$ given by:

$$\mathbb{P}\left(\mathbf{X}_t^{[-c_d^p, p]} = \tilde{\mathbf{k}} \mid \mathbf{X}_{t+1}^{[-c_d^p, p]} = \mathbf{x}_{t+1}^{[-c_d^p, p]}, \mathbf{Y}_{1:t}^{[-c_d^p, p]}, \tilde{\boldsymbol{\theta}}\right)$$
$$= \frac{\mathbb{P}\left(\mathbf{X}_{t+1}^{[-c_d^p, p]} = \mathbf{x}_{t+1}^{[-c_d^p, p]} \mid \mathbf{X}_t^{[-c_d^p, p]} = \tilde{\mathbf{k}}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_t^{[-c_d^p, p]} = \tilde{\mathbf{k}} \mid \mathbf{Y}_{1:t}^{[-c_d^p, p]}, \tilde{\boldsymbol{\theta}}\right)}{\sum\limits_{\tilde{\boldsymbol{\omega}} \in \mathcal{X}_s^{n_c^p - 1}} \mathbb{P}\left(\mathbf{X}_{t+1}^{[-c_d^p, p]} = \mathbf{x}_{t+1}^{[-c_d^p, p]} \mid \mathbf{X}_t^{[-c_d^p, p]} = \tilde{\boldsymbol{\omega}}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_t^{[-c_d^p, p]} = \tilde{\boldsymbol{\omega}} \mid \mathbf{Y}_{1:t}^{[-c_d^p, p]}, \tilde{\boldsymbol{\theta}}\right)},$$

and simulate a value for $\mathbf{X}_t^{[-c_d^p, p]}$ from the distribution defined by these probabilities.

7. For $t = T^{[c_d^p, p]}$ simulate a value for $\mathbf{X}_t^{[1:n_c^p, p]}$ from:

$$\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k} \mid \mathbf{X}_{t+1}^{[-c_d^p, p]} = \mathbf{x}_{t+1}^{[-c_d^p, p]}, \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \tilde{\boldsymbol{\theta}}\right)$$
$$= \frac{\mathbb{P}\left(\mathbf{X}_{t+1}^{[-c_d^p, p]} = \mathbf{x}_{t+1}^{[-c_d^p, p]} \mid \mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k} \mid \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \tilde{\boldsymbol{\theta}}\right)}{\sum\limits_{\boldsymbol{\omega} \in \mathcal{X}_s^{n_c^p}} \mathbb{P}\left(\mathbf{X}_{t+1}^{[-c_d^p, p]} = \mathbf{x}_{t+1}^{[-c_d^p, p]} \mid \mathbf{X}_t^{[1:n_c^p, p]} = \boldsymbol{\omega}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} = \boldsymbol{\omega} \mid \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \tilde{\boldsymbol{\theta}}\right)},$$

for $\mathbf{k} \in \mathcal{X}_s^{n_c^p}$.

8. Finally, for $t = T^{[c_d^p, p]} - 1, T^{[c_d^p, p]} - 2, \ldots, 1$ simulate a value for $\mathbf{X}_t^{[1:n_c^p, p]}$ from:

$$\mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k} \mid \mathbf{X}_{t+1}^{[1:n_c^p, p]} = \mathbf{x}_{t+1}^{[1:n_c^p, p]}, \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \tilde{\boldsymbol{\theta}}\right)$$
$$= \frac{\mathbb{P}\left(\mathbf{X}_{t+1}^{[1:n_c^p, p]} = \mathbf{x}_{t+1}^{[1:n_c^p, p]} \mid \mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} = \mathbf{k} \mid \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \tilde{\boldsymbol{\theta}}\right)}{\sum\limits_{\boldsymbol{\omega} \in \mathcal{X}_s^{n_c^p}} \mathbb{P}\left(\mathbf{X}_{t+1}^{[1:n_c^p, p]} = \mathbf{x}_{t+1}^{[1:n_c^p, p]} \mid \mathbf{X}_t^{[1:n_c^p, p]} = \boldsymbol{\omega}, \tilde{\boldsymbol{\phi}}\right) \mathbb{P}\left(\mathbf{X}_t^{[1:n_c^p, p]} = \boldsymbol{\omega} \mid \mathbf{Y}_{1:t}^{[1:n_c^p, p]}, \tilde{\boldsymbol{\theta}}\right)},$$

for $\mathbf{k} \in \mathcal{X}_s^{n_c^p}$.

B.3 Simulation studies: additional results

To simulate data with the same structure as the observations, we proceed as follows. We consider pen p in the field data ($p = 1, 2, \ldots, P$). We simulate an epidemic in a pen of size n_c^p from Equation 3.1 in Section 3.2 and obtain simulated hidden carriage state $X_t^{[c, p]}$ for each individual $c \in \{1, 2, \ldots, n_c^p\}$ at its observation period $t \in \mathcal{T}^{c,p} = \{1, 2, \ldots, T^{c,p}\}$. Conditional on these carriage states we then generate RAMS and faecal samples from individual c at the same days as those from individual c in pen

p of the real data, that is, at pre-specified observation times $O^{c,p}$. Eventually, the simulated data for pen p is the collection of $\left[\left(R_t^{[c,p]}, F_t^{[c,p]}\right)\right]_{c=1,2,\ldots,n_c^p;\, t\in O^{c,p}}$. The simulation of data is repeated for each pen.

Then we fit the basic HMM model described in Section 3.2. In Figure B.1 we show the median posterior medians along with 95% credible intervals over the 40 realisations of the two sampling schemes, under the 3 different scenarios.

B.4 Real data analysis of dataset 1: additional results

In this section we provide additional results for the analysis of the first *E. coli* O157:H7 data of Section 3.6.1. Trace plots of the model parameters are shown in Figure B.2, with the summaries of the marginal posterior distributions of the observation and transmission parameters given in Figure B.4 and B.3, respectively. Relying on the true test results for the RAMS and faecal test, the plot of the posterior probability of colonisation of each animal in a given pen is created. Two of the corresponding plots are displayed in Figures B.5 and B.6.

B.5 Real data analysis of dataset 2: additional results

In this section we provide additional results for the analysis of the second *E. coli* O157:H7 data of Section 3.6.2. Trace plots of the transmission parameters are given in Figure B.7. The posterior predictive distributions of the three test quantities are shown in Figure B.8. In Figure B.9 we plot the marginal posterior distributions of the parameters α, β, and m using four different choice of priors.

FIGURE B.1: Comparison of the distributions of the posterior median estimates of parameters based on 40 simulated data with different sampling scheme under three epidemic scenarios. The simulated data resemble the two motivating longitudinal studies. For each scenario the red dashed lines indicate the true values of the corresponding model parameter. Boxplots give the quantiles 2.5%, 25%, 50%, 75%, and 97.5%, respectively.

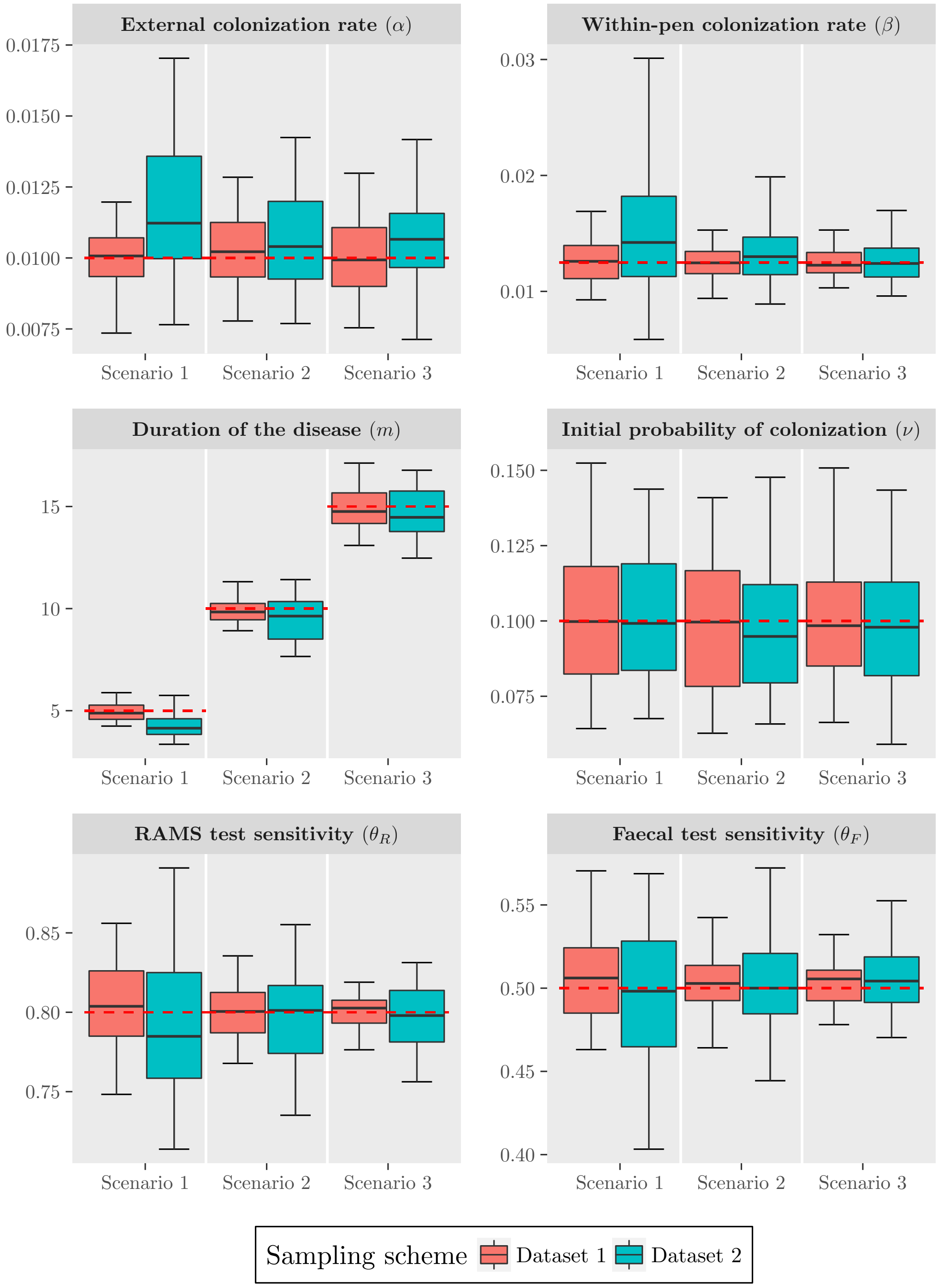


Figure B.2: Trace plot of each model parameter among cattle in the *E. coli* O157:H7 dataset 1.

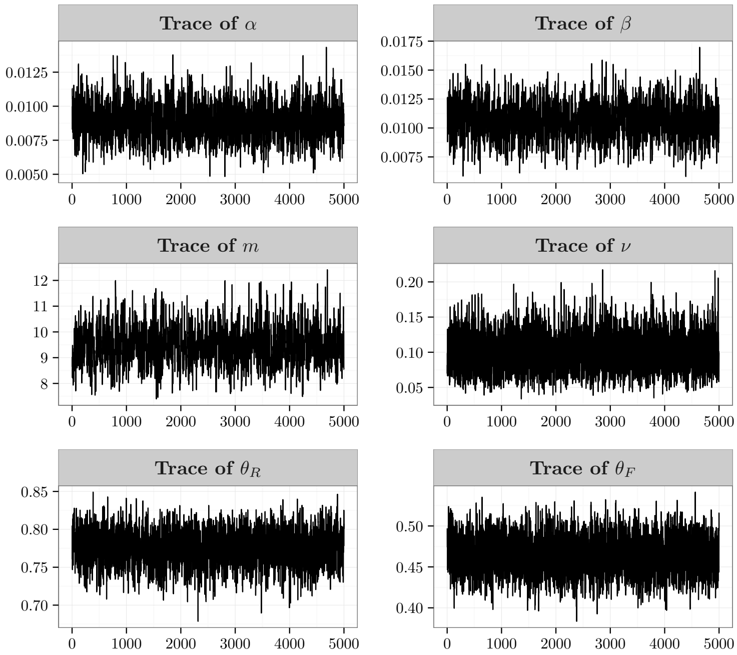


Figure B.3: Pairwise scatter-plots for each pair of observation parameters and marginal posterior densities (diagonal) among cattle in the *E. coli* O157:H7 dataset 1. Three independent Markov chains are presented.

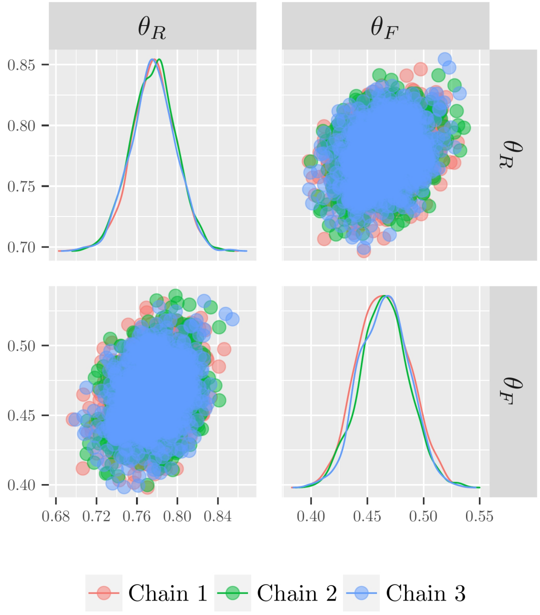


Figure B.4: Pairwise scatter-plots for each pair of transmission parameters and marginal posterior densities (diagonal) among cattle in the *E. coli* O157:H7 dataset 1. Three independent Markov chains are presented.

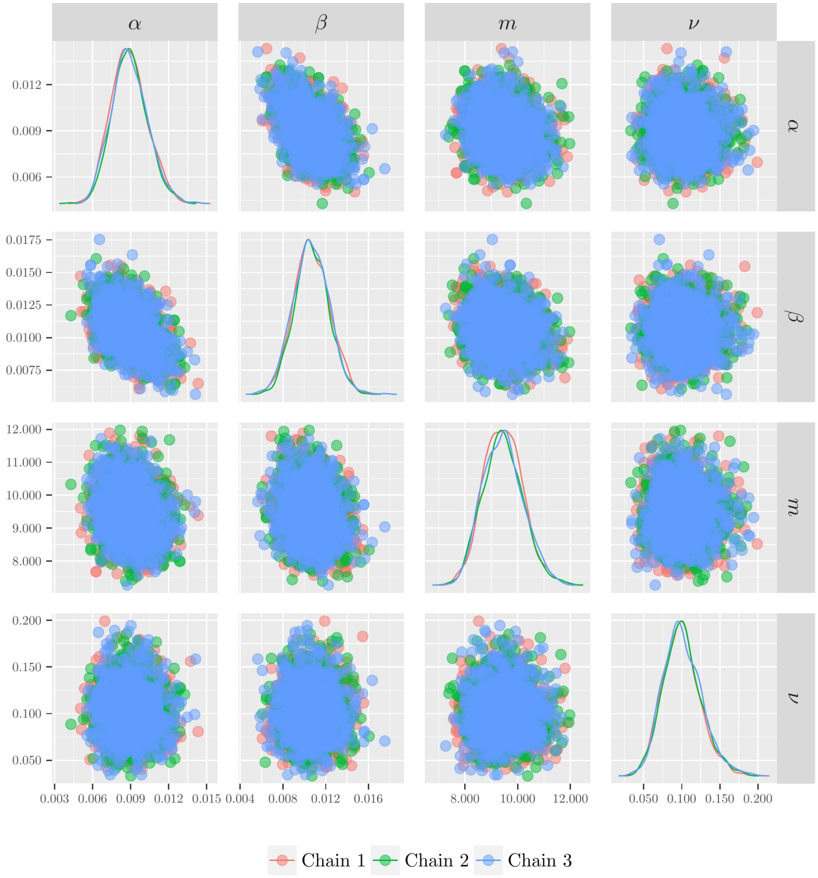


Figure B.5: Posterior probability of colonisation (grey solid line) for individuals in pen 3 of the first E. coli O157:H7 dataset, over the entire sampling period of 99 days (1 for colonised, 0 for non-carrier). For reference we also show test results taken twice per week; "·" indicates negative sample and "+" indicates that the sample was positive. White horizontal lines represent the days in which samples were taken.

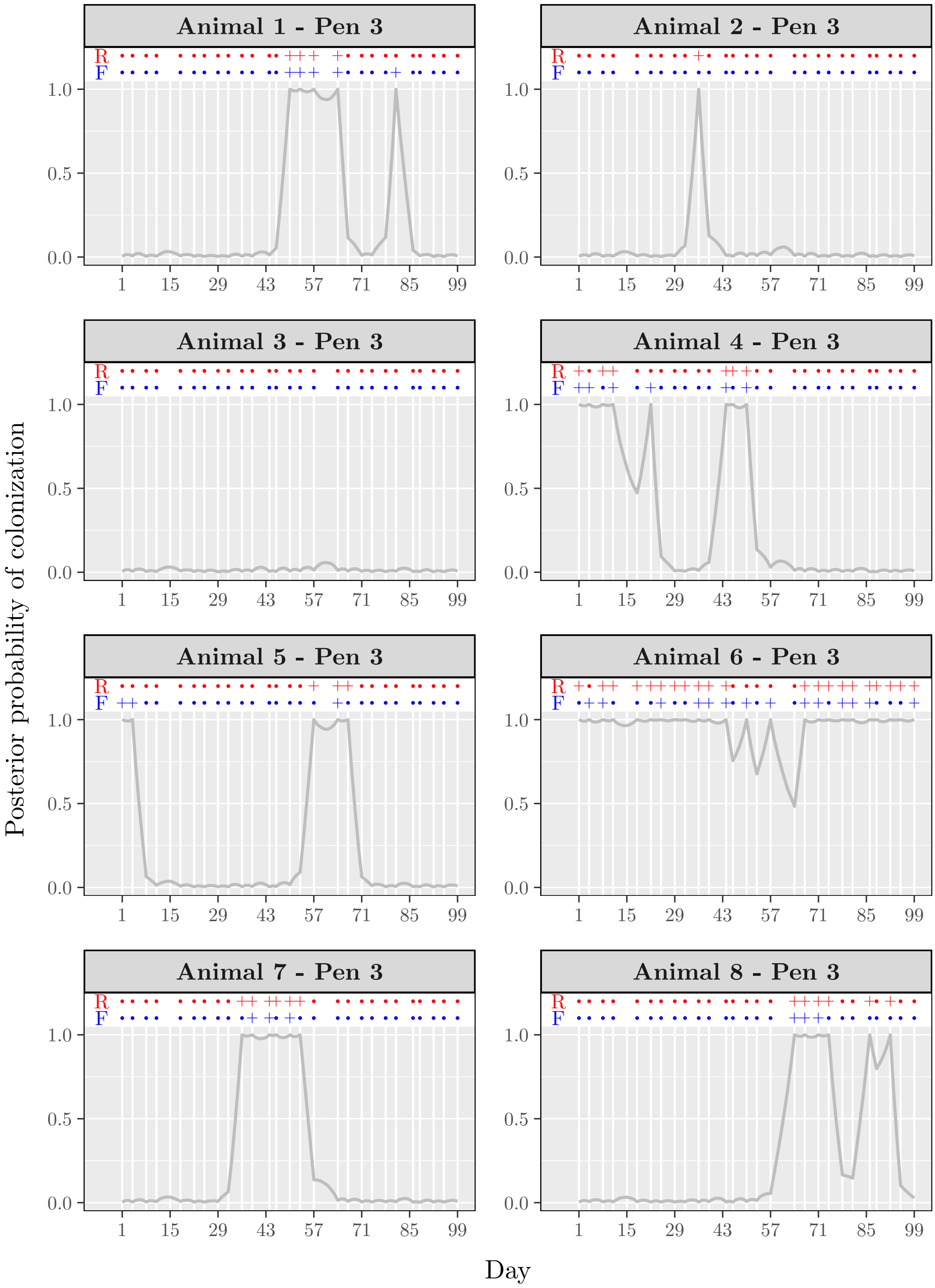


Figure B.6: Posterior probability of colonisation (grey solid line) for individuals in pen 19 of the first E. coli O157:H7 dataset, over the entire sampling period of 99 days (1 for colonised, 0 for non-carrier). For reference we also show test results taken twice per week; ”·” indicates negative sample and ”+” indicates that the sample was positive. White horizontal lines represent the days in which samples were taken.

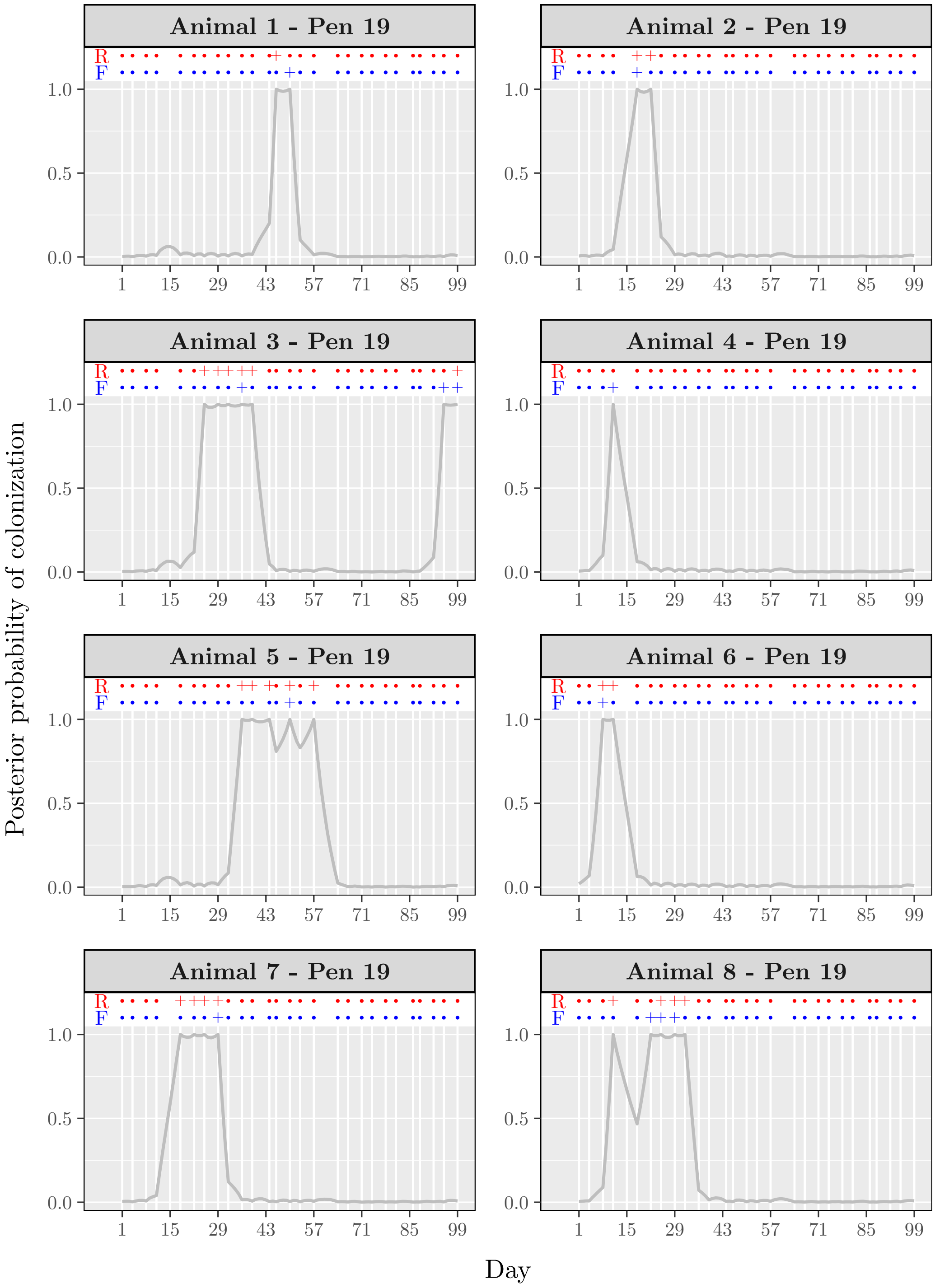


Figure B.7: Trace plot of each model parameter among cattle in the *E. coli* O157:H7 dataset 2. Three independent Markov chains are presented.

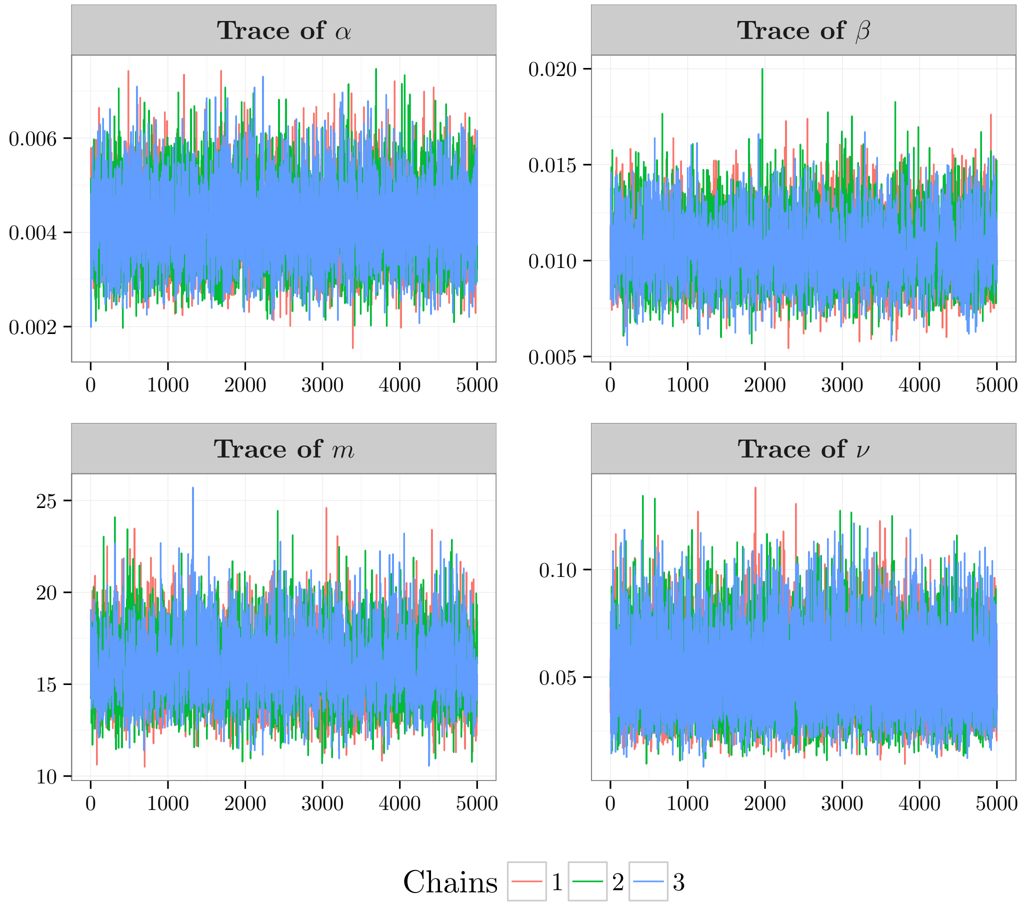


Figure B.8: Model assessment plots for *E. coli* O157:H7 dataset 2. Posterior predictive distribution of the mean duration, number of animals that never tested as positive and total numbers of positive test results. Black dashed line indicate the observed value of the corresponding summary. Shaded area corresponds to the 95% credible interval. The results are based on 5000 posterior predictive simulations having the same structure as in the original dataset.

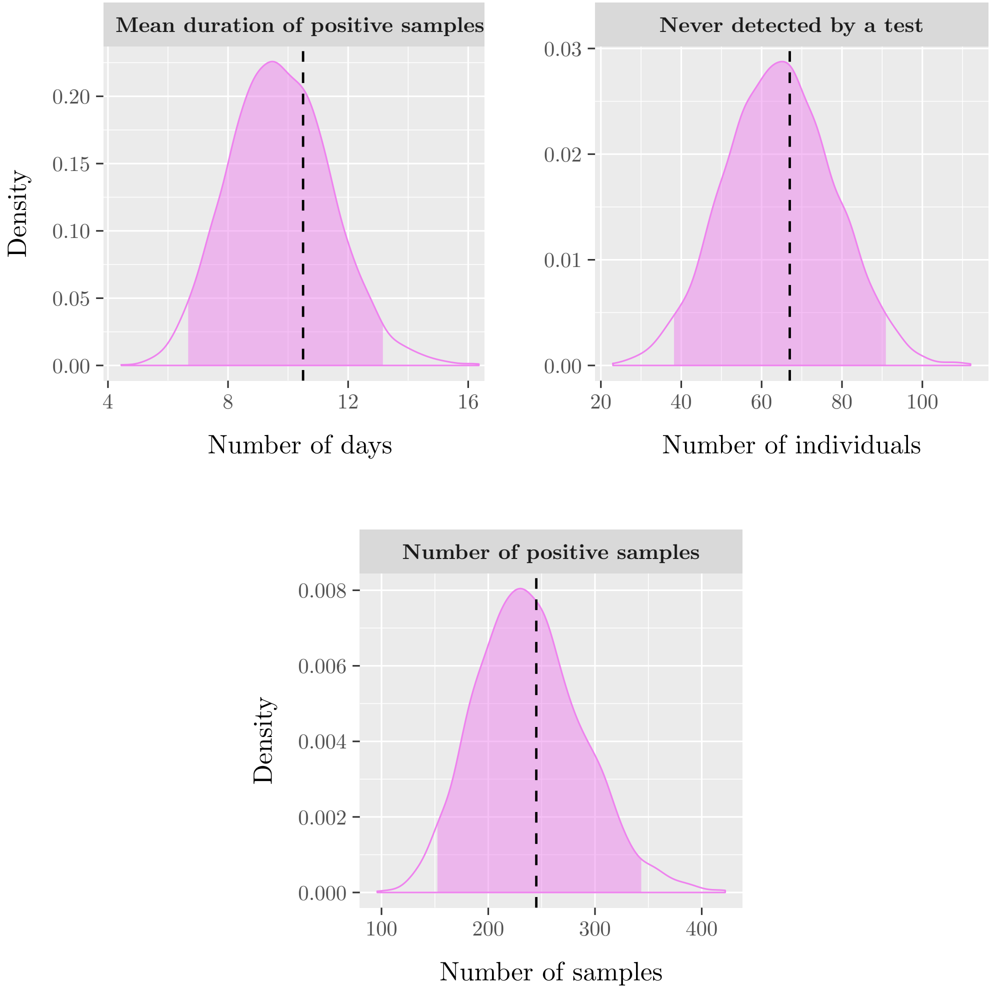
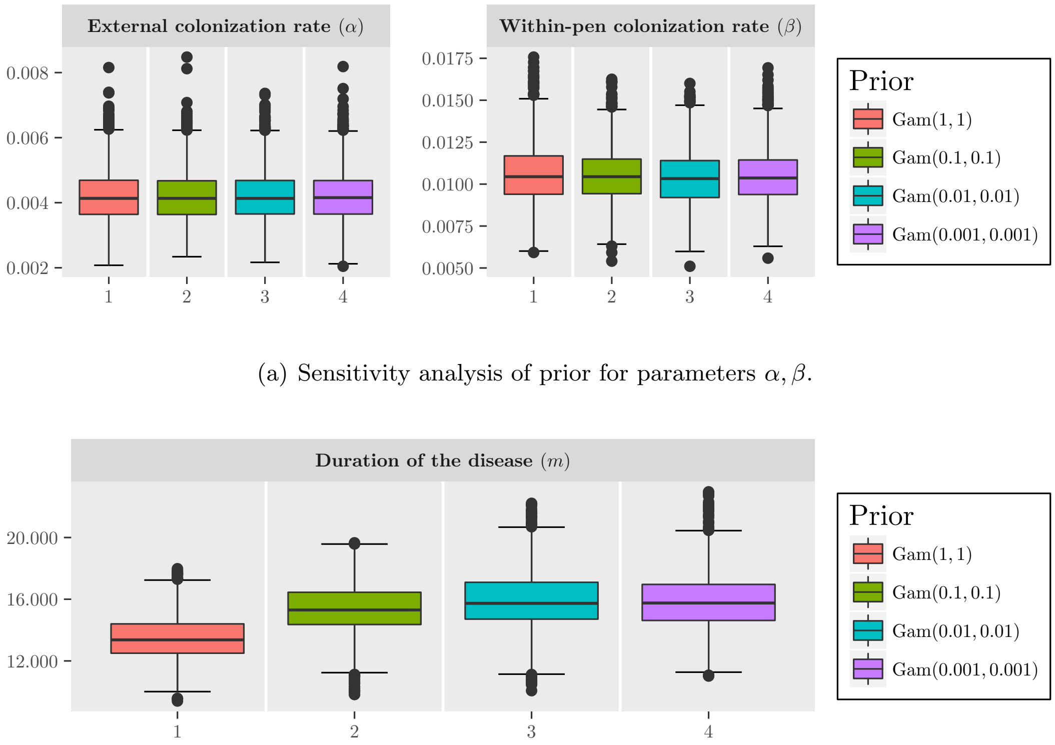


Figure B.9: Sensitivity to the prior distribution on the colonisation rates and the mean colonisation duration for *E. coli* O157:H7 dataset 2. We use 4 different Gamma priors, 1-4, with constant mean equal to 1 and variance equal to 1, 10, 100 and 1000 respectively. Each time, the prior of only one parameter is changed.



(a) Sensitivity analysis of prior for parameters α, β.

(b) Sensitivity analysis of prior for parameters m.

Appendix C

Efficient Model Comparison Techniques Supplementary Material

C.1 Algorithm details

In this section we compare the variance of the estimators $\widehat{P}_{rq}$ and $\widetilde{P}_{rq}$. For $M = 1, 2, \ldots$, let:

$$\tilde{P}_{rq}^{M} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{\pi\left(\mathbf{y} \mid \mathbf{x}^{(i,j)}, \boldsymbol{\theta}^{(i)}\right) \pi\left(\mathbf{x}^{(i,j)} \mid \boldsymbol{\theta}^{(i)}\right) \pi\left(\boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x}^{(i,j)} \mid \boldsymbol{\theta}^{(i)}\right) q\left(\boldsymbol{\theta}^{(i)}\right)}, \tag{C.1}$$

where $\boldsymbol{\theta}^{(i)} \sim q(\cdot)$ and $\mathbf{x}^{(i,j)} \sim r\left(\cdot \mid \boldsymbol{\theta}^{(i)}\right)$. Thus $\widehat{P}_{rq}$ is the special case $\tilde{P}_{rq}^{1}$.

To show that Equation (C.1) is unbiased we can consider the following:

$$\mathbb{E}[\tilde{P}_{rq}^{M}] = \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}}\left[\tilde{P}_{rq}^{M} \mid \boldsymbol{\theta}\right]\right], \tag{C.2}$$

by the law of total expectation. Hence,

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}}\left[\tilde{P}_{rq}^{M} \mid \boldsymbol{\theta}\right] &= \frac{1}{NM} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}}\left[\frac{\pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)} \sum_{j=1}^{M} \frac{\pi\left(\mathbf{y} \mid \mathbf{x}^{(i,j)}, \boldsymbol{\theta}^{(i)}\right) \pi\left(\mathbf{x}^{(i,j)} \mid \boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x}^{(i,j)} \mid \boldsymbol{\theta}^{(i)}\right)}\right] \\
&= \frac{1}{NM} \sum_{i=1}^{N} \left[\frac{\pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)} \sum_{j=1}^{M} \mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}}\left[\frac{\pi\left(\mathbf{y} \mid \mathbf{x}^{(i,j)}, \boldsymbol{\theta}^{(i)}\right) \pi\left(\mathbf{x}^{(i,j)} \mid \boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x}^{(i,j)} \mid \boldsymbol{\theta}^{(i)}\right)}\right]\right] \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{\pi\left(\mathbf{y} \mid \boldsymbol{\theta}^{(i)}\right) \pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)},
\end{aligned} \tag{C.3}$$

and therefore from Equation (C.2) and Equation (C.3) we have:

$$\begin{aligned}
\mathbb{E}\left[\tilde{P}_{rq}^{M}\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{\mathbf{x}\mid\boldsymbol{\theta}}\left[\tilde{P}_{rq}^{M}\mid\boldsymbol{\theta}\right]\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\pi\left(\mathbf{y}\mid\boldsymbol{\theta}^{(i)}\right)\pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\pi\left(\mathbf{y}\mid\boldsymbol{\theta}^{(i)}\right)\pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\pi(\mathbf{y}) \\
&= \pi(\mathbf{y}). \qquad \text{(C.4)}
\end{aligned}$$

To calculate the variance of Equation (C.1), we can consider the law of total variance:

$$\operatorname{Var}\left(\tilde{P}_{rq}^{M}\right) = \mathbb{E}_{\boldsymbol{\theta}}\left[\operatorname{Var}_{\mathbf{x}\mid\boldsymbol{\theta}}\left(\tilde{P}_{rq}^{M}\mid\boldsymbol{\theta}\right)\right] + \operatorname{Var}_{\boldsymbol{\theta}}\left[\mathbb{E}_{\mathbf{x}\mid\boldsymbol{\theta}}\left(\tilde{P}_{rq}^{M}\mid\boldsymbol{\theta}\right)\right]. \qquad \text{(C.5)}$$

To evaluate Equation (C.5) we need to consider:

$$\begin{aligned}
\operatorname{Var}_{\mathbf{x}\mid\boldsymbol{\theta}}\left(\tilde{P}_{rq}^{M}\mid\boldsymbol{\theta}\right) &= \frac{1}{N^2}\sum_{i=1}^{N}\operatorname{Var}_{\mathbf{x}\mid\boldsymbol{\theta}}\left[\frac{\pi\left(\boldsymbol{\theta}^{(i)}\right)}{Mq\left(\boldsymbol{\theta}^{(i)}\right)}\sum_{j=1}^{M}\frac{\pi\left(\mathbf{y}\mid\mathbf{x}^{(i,j)},\boldsymbol{\theta}^{(i)}\right)\pi\left(\mathbf{x}^{(i,j)}\mid\boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x}^{(i,j)}\mid\boldsymbol{\theta}^{(i)}\right)}\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{\pi\left(\boldsymbol{\theta}^{(i)}\right)}{Mq\left(\boldsymbol{\theta}^{(i)}\right)}\right)^2\sum_{j=1}^{M}\operatorname{Var}_{\mathbf{x}\mid\boldsymbol{\theta}}\left(\frac{\pi\left(\mathbf{y}\mid\mathbf{x}^{(i,j)},\boldsymbol{\theta}^{(i)}\right)\pi\left(\mathbf{x}^{(i,j)}\mid\boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x}^{(i,j)}\mid\boldsymbol{\theta}^{(i)}\right)}\right). \qquad \text{(C.6)}
\end{aligned}$$

In order to calculate Equation (C.6) we need to consider:

$$\begin{aligned}
&\operatorname{Var}_{\mathbf{x}\mid\boldsymbol{\theta}}\left(\frac{\pi\left(\mathbf{y}\mid\mathbf{x}^{(i,j)},\boldsymbol{\theta}^{(i)}\right)\pi\left(\mathbf{x}^{(i,j)}\mid\boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x}^{(i,j)}\mid\boldsymbol{\theta}^{(i)}\right)}\right) \\
&\qquad = \int_{\mathbf{x}}\left(\frac{\pi\left(\mathbf{y}\mid\mathbf{x},\boldsymbol{\theta}^{(i)}\right)\pi\left(\mathbf{x}\mid\boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x}\mid\boldsymbol{\theta}^{(i)}\right)} - \pi\left(x\mid\boldsymbol{\theta}^{(i)}\right)\right)^2 \times r\left(\mathbf{x}\mid\boldsymbol{\theta}^{(i)}\right)\mathrm{d}\mathbf{x} \\
&\qquad = \pi\left(\mathbf{y}\mid\boldsymbol{\theta}^{(i)}\right)^2\int_{\mathbf{x}}\left(\frac{\pi\left(\mathbf{x}\mid\mathbf{y},\boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x}\mid\boldsymbol{\theta}^{(i)}\right)} - 1\right)^2 r\left(\mathbf{x}\mid\boldsymbol{\theta}^{(i)}\right)\mathrm{d}\mathbf{x}, \qquad \text{(C.7)}
\end{aligned}$$

which is constant for a given $\boldsymbol{\theta}^{(i)}$. Hence, if we let:

$$v\left(\boldsymbol{\theta}^{(i)}\right) = \int_{\mathbf{x}} \left(\frac{\pi\left(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}^{(i)}\right)}{r\left(\mathbf{x} \mid \boldsymbol{\theta}^{(i)}\right)} - 1 \right)^2 r\left(\mathbf{x} \mid \boldsymbol{\theta}^{(i)}\right) \mathrm{d}\mathbf{x},$$

derived in Equation (C.7), then from Equation (C.6) we have:

$$\begin{aligned} \mathrm{Var}_{\mathbf{x}|\boldsymbol{\theta}}\left(\tilde{P}_{rq}^{M} \mid \boldsymbol{\theta}\right) &= \frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{\pi\left(\mathbf{y} \mid \boldsymbol{\theta}^{(i)}\right)\pi\left(\boldsymbol{\theta}^{(i)}\right)}{Mq\left(\boldsymbol{\theta}^{(i)}\right)}\right)^2 \sum_{j=1}^{M} v\left(\boldsymbol{\theta}^{(i)}\right) \\ &= \frac{1}{MN^2}\sum_{i=1}^{N}\left(\frac{\pi\left(\mathbf{y} \mid \boldsymbol{\theta}^{(i)}\right)\pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)}\right)^2 v\left(\boldsymbol{\theta}^{(i)}\right), \end{aligned} \tag{C.8}$$

and from this we can calculate:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{Var}_{\mathbf{x}|\boldsymbol{\theta}}\left(\tilde{P}_{rq}^{M} \mid \boldsymbol{\theta}\right)\right] &= \frac{1}{MN^2}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\pi\left(\mathbf{y} \mid \boldsymbol{\theta}^{(i)}\right)\pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)}\right)^2 v\left(\boldsymbol{\theta}^{(i)}\right)\right] \\ &= \frac{1}{MN^2}\sum_{i=1}^{N}\int_{\boldsymbol{\theta}}\frac{\left[\pi\left(\mathbf{y} \mid \boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)\right]^2 v\left(\boldsymbol{\theta}\right)}{q\left(\boldsymbol{\theta}\right)}\mathrm{d}\boldsymbol{\theta} \\ &= \frac{\pi\left(\mathbf{y}\right)^2}{MN^2}\sum_{i=1}^{N}\int_{\boldsymbol{\theta}}\frac{\pi\left(\boldsymbol{\theta} \mid \mathbf{y}\right)^2 v\left(\boldsymbol{\theta}\right)}{q\left(\boldsymbol{\theta}\right)}\mathrm{d}\boldsymbol{\theta} \\ &= \frac{\pi\left(\mathbf{y}\right)^2}{MN}v_2\left(\mathbf{y}\right), \end{aligned} \tag{C.9}$$

where $v_2\left(\mathbf{y}\right) = \int_{\boldsymbol{\theta}} \frac{\pi(\boldsymbol{\theta}|\mathbf{y})^2 v(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}$.

The final component of Equation (C.5) we need to calculate is:

$$\begin{aligned} \mathrm{Var}_{\boldsymbol{\theta}}\left[\mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}}\left(\tilde{P}_{rq}^{M} \mid \boldsymbol{\theta}\right)\right] &= \mathrm{Var}_{\boldsymbol{\theta}}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\pi\left(\mathbf{y} \mid \boldsymbol{\theta}^{(i)}\right)\pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)}\right] \qquad \text{from (C.3)} \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}_{\boldsymbol{\theta}}\left(\frac{\pi\left(\mathbf{y} \mid \boldsymbol{\theta}^{(i)}\right)\pi\left(\boldsymbol{\theta}^{(i)}\right)}{q\left(\boldsymbol{\theta}^{(i)}\right)}\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\int_{\boldsymbol{\theta}}\left(\frac{\pi\left(\mathbf{y} \mid \boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{q\left(\boldsymbol{\theta}\right)} - \pi(\mathbf{y})\right)^2 q\left(\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{\theta} \\ &= \frac{\pi(\mathbf{y})^2}{N^2}\sum_{i=1}^{N}\int_{\boldsymbol{\theta}}\left(\frac{\pi\left(\boldsymbol{\theta} \mid \mathbf{y}\right)}{q\left(\boldsymbol{\theta}\right)} - 1\right)^2 q\left(\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{\theta} \end{aligned}$$

$$= \frac{\pi(\mathbf{y})^2}{N} v_3(\mathbf{y}), \tag{C.10}$$

where $v_3(\mathbf{y}) = \int_{\boldsymbol{\theta}} \left(\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})} - 1 \right)^2 q\left(\boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\theta}$.

Finally, from Equations (C.5), (C.9) and (C.10) we have:

$$\begin{aligned}
\operatorname{Var}\left(\tilde{P}_{rq}^{M}\right) &= \frac{\pi\left(\mathbf{y}\right)^2}{MN} v_2\left(\mathbf{y}\right) + \frac{\pi(\mathbf{y})^2}{N} v_3(\mathbf{y}) \\
&= \frac{\pi\left(\mathbf{y}\right)^2}{N} \left(\frac{v_2\left(\mathbf{y}\right)}{M} + v_3(\mathbf{y}) \right). \qquad \text{(C.11)}
\end{aligned}$$

It is clear from the above that for fixed NM, choosing $M = 1$ minimises $\operatorname{Var}\left(\tilde{P}_{rq}^{M}\right)$. Note that the computational cost of computing $\operatorname{Var}\left(\tilde{P}_{rq}^{M}\right)$ is not the same for all NM as N $\boldsymbol{\theta}$ samples are drawn from $q(\cdot)$ and for each $i = 1, 2, \ldots, N$, M samples $\mathbf{x}^{(i,\cdot)}$ from $r\left(\cdot \mid \boldsymbol{\theta}^{(i)}\right)$ are made. However sampling $\boldsymbol{\theta}$ generally takes negligible time compared to sampling $\mathbf{x}$ and thus throughout the thesis we take $M = 1$.

C.2 Implementation details

In this section we briefly overview alternative techniques for estimating the likelihood or the posterior probabilities in the presence of missing data. In particular, we explain the adaptations that are required in order to implement the methods described in Section 1.2.4.

C.2.1 Marginal likelihood estimation via harmonic mean

When data augmentation is used, the parameter vector comprises latent variable $\mathbf{x}$ as well as the model parameters $\boldsymbol{\theta}$. The marginal likelihood $\pi(\mathbf{y})$ can be approximated by the sample harmonic mean of the likelihoods,

$$\hat{P}_{HM}(\mathbf{y}) = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{1}{\pi\left(\mathbf{y} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(i)}\right)} \right]^{-1} \tag{C.12}$$

based on N MCMC draws $\left(\mathbf{x}^{(1)}, \boldsymbol{\theta}^{(1)}\right), \left(\mathbf{x}^{(2)}, \boldsymbol{\theta}^{(2)}\right), \ldots, \left(\mathbf{x}^{(N)}, \boldsymbol{\theta}^{(N)}\right)$ from the joint posterior $\pi(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})$.

C.2.2 Marginal likelihood estimation via bridge sampling

With missing data the bridge sampling estimator can be written as,

$$\widehat{P}_{\mathrm{BS}}^{(t)}(\mathbf{y}) = \widehat{P}_{\mathrm{BS}}^{(t-1)}(\mathbf{y}) \, \frac{\displaystyle \frac{1}{L}\sum_{\ell=1}^{L} \frac{1}{\dfrac{L\, q\left(\tilde{\boldsymbol{\theta}}^{(\ell)}\right)\, \pi\left(\tilde{\mathbf{x}}^{(\ell)} \mid \mathbf{y}, \tilde{\boldsymbol{\theta}}^{(\ell)}\right) \widehat{P}_{\mathrm{BS}}^{(t-1)}(\mathbf{y})}{\pi\left(\mathbf{y}, \tilde{\mathbf{x}}^{(\ell)} \mid \tilde{\boldsymbol{\theta}}^{(\ell)}\right) \pi\left(\tilde{\boldsymbol{\theta}}^{(\ell)}\right)} + N}}{\displaystyle \frac{1}{N}\sum_{i=1}^{N} \frac{1}{L + \dfrac{N\, \pi\left(\mathbf{y}, \hat{\mathbf{x}}^{(i)} \mid \hat{\boldsymbol{\theta}}^{(i)}\right) \pi\left(\hat{\boldsymbol{\theta}}^{(i)}\right)}{q\left(\hat{\boldsymbol{\theta}}^{(i)}\right)\, \pi\left(\hat{\mathbf{x}}^{(i)} \mid \mathbf{y}, \hat{\boldsymbol{\theta}}^{(i)}\right) \widehat{P}_{\mathrm{BS}}^{(t-1)}(\mathbf{y})}}}, \qquad \text{(C.13)}$$

using MCMC iid draws $\left(\hat{\mathbf{x}}^{(i)}, \hat{\boldsymbol{\theta}}^{(i)}\right)$, $i = 1, 2, \ldots, N$ from the joint posterior $\pi(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})$ and $\tilde{\boldsymbol{\theta}}^{(\ell)}$, $\ell = 1, 2, \ldots, L$ from the importance sampling density $q(\boldsymbol{\theta})$. For each sample $\tilde{\boldsymbol{\theta}}^{(\ell)}$ we obtain a corresponding sample for the missing data $\tilde{\mathbf{x}}^{(\ell)}$ from $\pi\left(\mathbf{x} \mid \mathbf{y}, \tilde{\boldsymbol{\theta}}^{(\ell)}\right)$. The proposed importance sampling estimator, which is a special case of the general bridge sampling estimator, is used as starting value $\widehat{P}_{\mathrm{BS}}^{(0)}(\mathbf{y})$. In practise, iterative application of Equation (C.13) is very fast.

C.2.3 Marginal likelihood estimation via Chib's method

Chib's method for estimating the marginal likelihood can be calculated by:

$$\log \widehat{P}_{\mathrm{Chib}}(\mathbf{y}) = \log \pi(\mathbf{y}, \mathbf{x}^* \mid \boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \log \widehat{\pi}(\mathbf{x}^*, \boldsymbol{\theta}^* \mid \mathbf{y}). \qquad \text{(C.14)}$$

For the model described in Section 4.4.1 we decompose the parameter vector into $(\mathbf{x}, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$, where $\mathbf{b}_1 = (\alpha_1, \alpha_2, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \mu_1, \mu_2, w)$, $\mathbf{b}_2 = \nu_1$ and $\mathbf{b}_3 = \nu_2$. The posterior density is then factorised as:

$$\pi(\mathbf{x}^*, \boldsymbol{\theta}^* \mid \mathbf{y}) = \pi(\mathbf{x}^* \mid \mathbf{y})\, \pi(\mathbf{b}_1^* \mid \mathbf{y}, \mathbf{x}^*)\, \pi(\mathbf{b}_2^* \mid \mathbf{y}, \mathbf{x}^*, \mathbf{b}_1^*)\, \pi(\mathbf{b}_3^* \mid \mathbf{y}, \mathbf{x}^*, \mathbf{b}_1^*, \mathbf{b}_2^*).$$

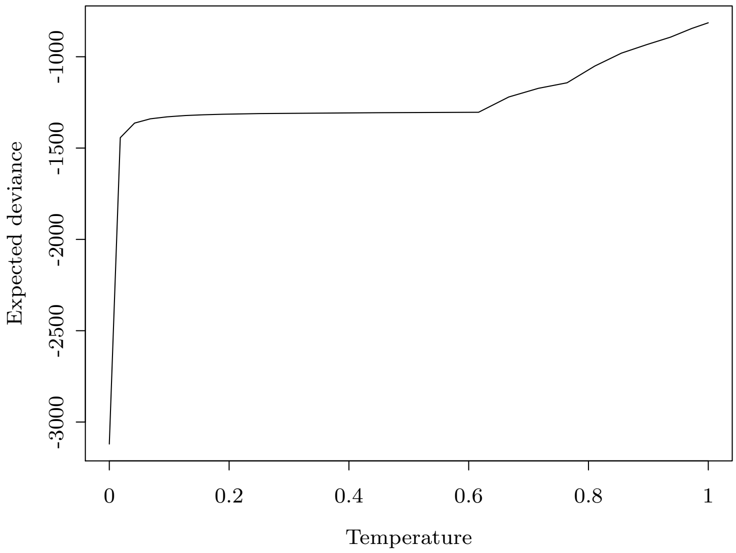
C.2.4 Marginal likelihood estimation via power posteriors

The power posterior approach to estimating the marginal likelihood, as described in Section 1.2.4.6, can be implement once we include the missing data in the set of unknown parameters. Metropolis within Gibbs sampling was used to obtain samples from the power posterior $\pi(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y}, t)$ at each temperature $t > 0$.

The variability of the power posterior estimator depends on the chosen number and spacing of the t_i's. Choosing a large number of temperatures, the estimation of the log marginal likelihood requires considerably more computational effort. Moreover, the precision of the estimate is sensitive to the number of samples used and the mixing of the MCMC sampler.

In Friel and Pettitt (2008) the temperatures were chosen with a geometric spacing, $t_l = (l/n)^c$, for $l = 0, 1, \ldots, n$, with $c > 1$, which places many of the temperatures close to zero. This scheme is preferable in cases where the expected deviance has a sharp increase near zero before leveling off. However, in our case, the curve of the expected deviance is not convex (Figure C.1). After some pilot analysis (not counted in the computation cost) using the setup that we described in Section 4.4.1 we chose to use 20 partitions of the unit line, placing more temperatures around zero and the other sharp change.

Figure C.1: Expected deviance against temperature for model $\mathcal{M}_1$ estimated using power posteriors.



C.2.5 Reversible jump MCMC

In this section, we provide details of the reversible jump algorithm adjusted to compare model $\mathcal{M}_1$ (described in Section 4.3.1) with the nested model $\mathcal{M}_2$, in which the community acquisition rates for adults and children are equal (Section 4.4.2.1). The main difficulty with RJMCMC lies in designing efficient proposals to jump between models and their associated parameters.

More specifically, when the algorithm is in model $\mathcal{M}_1$, we propose a move to

$\mathcal{M}_2$ with probability 0.5, in which the joint community acquisition rate α is set to $\alpha = \frac{L_1\,\alpha_1 + L_2\,\alpha_2}{L_1 + L_2}$, where L_1 is the total number of children and L_2 is the total number of adults. The Jacobian of the transformation is $\frac{L_1\,L_2}{L_1+L_2}$. For the reverse move, we need to increase the dimension of the parameter vector, therefore an auxiliary random variable u is required. Let $u \sim \mathcal{N}(0, \sigma^2)$ with σ^2 fixed but well chosen. Then we set $\alpha_1 = \alpha + \frac{u}{L_1}$ and $\alpha_2 = \alpha - \frac{u}{L_2}$. The Jacobian of the transformation is then $\frac{L_1+L_2}{L_1\,L_2}$. The acceptance probability of jumping from $\mathcal{M}_1$ to $\mathcal{M}_2$, is given by $\min(1, A_{12})$ where,

$$A_{12} = \frac{\pi(k=2, \boldsymbol{\psi}_2 \mid \mathbf{y})\, \pi(k=2)}{\pi(k=1, \boldsymbol{\psi}_1 \mid \mathbf{y})\, \pi(k=1)} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(\frac{L_1 L_2(\alpha_1-\alpha_2)}{L_1+L_2}\right)^2} \right) \frac{L_1\,L_2}{L_1 + L_2},$$

where $\boldsymbol{\psi}_1 = (\alpha_1, \alpha_2, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, w, \mu_1, \mu_2, \nu_1, \nu_2, \mathbf{x})$ and $\boldsymbol{\psi}_2 = (\alpha, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, w, \mu_1, \mu_2, \nu_1, \nu_2, \mathbf{x})$. For the reciprocal move from model $\mathcal{M}_2$ to $\mathcal{M}_1$, the probability of accepting the jump is given by $\min(1, A_{21})$ where,

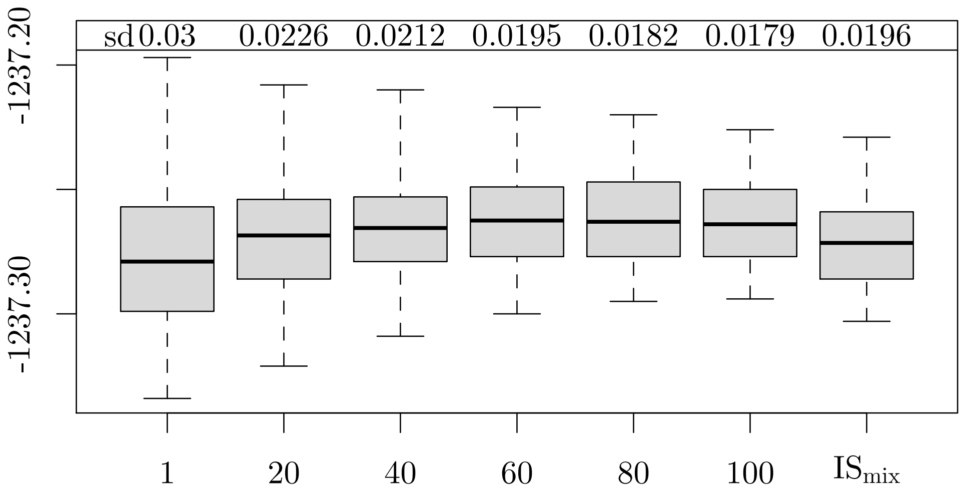
$$A_{21} = \frac{\pi(k=1, \boldsymbol{\psi}_1 \mid \mathbf{y})\, \pi(k=1)}{\pi(k=2, \boldsymbol{\psi}_2 \mid \mathbf{y})\, \pi(k=2)} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma^2}} \right)^{-1} \frac{L_1 + L_2}{L_1\,L_2}.$$

In addition to the model-switching step, the within-model parameters are updated using a standard MCMC algorithm that employs both Gibbs sampler updates and random walk Metropolis steps with a Gaussian proposal density centred at the current value.

C.3 Thinning for the bridge sampling

The optimal bridge sampling estimator is constructed on the basis of having iid samples from the posterior available. Thinning can be used to reduce the autocorrelation in posterior samples produced using MCMC. Figure C.2 shows the effect that the amount of thinning has on the bridge sampling estimator. Interestingly, quite substantial thinning is needed before the Monte Carlo variance of the bridge sampling estimator drops below that of the importance sampling estimator.

Figure C.2: Effect of thinning of the MCMC samples on the Monte Carlo variance of the bridge sampling estimator, over 50 replicates.



Appendix D

Evaluation Of Epidemiological Hypotheses Supplementary Material

D.1 HMC details for updating the transmission parameters of the Negative Binomial model

In this section we provide the details of the HMC algorithm, used to generate samples from the posterior of the Negative Binomial transmission model presented in Section 5.2. The joint posterior distribution of the hidden colonisation states and the parameters is given by,

$$\pi(\mathbf{X}, \tilde{\alpha}, \tilde{\beta}, \tilde{m}, \kappa, \nu, \theta_R, \theta_F \mid \mathbf{Y}) \propto \pi(\mathbf{Y} \mid \mathbf{X}, \theta_R, \theta_F)\ \pi(\mathbf{X} \mid \tilde{\alpha}, \tilde{\beta}, \tilde{m}, \kappa, \nu)\, \pi(\tilde{\boldsymbol{\theta}}), \quad \text{(D.1)}$$

where $\pi(\tilde{\boldsymbol{\theta}})$ is the prior of the transformed model parameters $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\phi}}, \boldsymbol{\vartheta}) = (\tilde{\alpha}, \tilde{\beta}, \tilde{m}, \kappa, \nu, \theta_R, \theta_F)$, $\tilde{\alpha} = \log(\alpha)$, $\tilde{\beta} = \log(\beta)$ and $\tilde{m} = m - 1$. The prior distributions are specified exactly as in the Geometric model for the parameters $\alpha, \beta, \tilde{m}, \nu, \theta_R$ and θ_F. We choose a Gamma prior for the additional parameter $\kappa \sim \text{Ga}(b_\kappa, c_\kappa)$.

The first term in Equation (D.1) can be written as,

$$\pi(\mathbf{Y} \mid \mathbf{X}, \theta_R, \theta_F) = \prod_{p=1}^{P} \prod_{c=1}^{n_c^p} \prod_{t=1}^{T^p} f_{x_t^{[c,\,p]}} \left(y_t^{[c,\,p]} \mid \theta_R, \theta_F \right),$$

where $f_{x_t^{[c,\,p]}} \left(y_t^{[c,\,p]} \mid \theta_R, \theta_F \right)$ is given by Equation (3.4), and the second term is given by,

$$\pi(\mathbf{X} \mid \tilde{\alpha}, \tilde{\beta}, \tilde{m}, \kappa, \nu) = \prod_{p=1}^{P} \prod_{t=2}^{T^p} \left[\left(1 - \exp\left\{ - e^{\tilde{\alpha}} - e^{\tilde{\beta}} \sum_{c=1}^{n_c^p} x_{t-1}^{[c,\,p]} \right\} \right)^{N_{01}^p(t)} \right.$$

$$\times \left(\exp\left\{ - e^{\tilde{\alpha}} - e^{\tilde{\beta}} \sum_{c=1}^{n_c^p} x_{t-1}^{[c,\,p]} \right\} \right)^{N_{00}^p(t)} \Bigg] \times \prod_{p=1}^{P} \prod_{c=1}^{n_c^p} \left[\nu^{x_1^{[c,\,p]}} (1-\nu)^{1-x_1^{[c,\,p]}} \right]$$

$$\times \prod_{p=1}^{P} \prod_{c=1}^{n_c^p} \left[\left[\frac{1 - \sum\limits_{\zeta=1}^{\zeta_1^{*\,[c,\,p]}-1} \mathbb{P}\big(Z=\zeta\big)}{\tilde{m}+1} \right]^{X_1^{[c,\,p]}} \left[\mathbb{P}\big(Z = \zeta_1^{*\,[c,\,p]}\big) \right]^{1-X_1^{[c,\,p]}} \right.$$

$$\left. \times \left[1 - \sum_{\zeta=1}^{\zeta^{*\,[c,\,p]}_{\tau^{[c,\,p]}}-1} \mathbb{P}\big(Z=\zeta\big) \right]^{X^{[c,\,p]}_{T^{[c,\,p]}}} \left[\mathbb{P}\big(Z = \zeta^{*\,[c,\,p]}_{\tau^{[c,\,p]}}\big) \right]^{1-X^{[c,\,p]}_{T^{[c,\,p]}}} \prod_{t=2}^{\tau^{[c,\,p]}-1} \mathbb{P}\big(Z = \zeta_t^{*\,[c,\,p]}\big) \right],$$

where $N_{0j}^p(t)$ denotes the number of individuals in pen p who were in state 0 at time $t-1$ and in state j at time t, for $j \in \{0,1\}$, $\Big(\zeta_1^{*\,[c,\,p]}, \zeta_2^{*\,[c,\,p]}, \ldots, \zeta^{*\,[c,\,p]}_{\tau^{[c,\,p]}}\Big)$ denotes the observed colonisation durations vector for each individual c in pen p, and

$$\mathbb{P}(Z=\zeta) = \left(\frac{\kappa}{\kappa+\tilde{m}}\right)^{\kappa} \frac{\Gamma(\kappa+\zeta-1)}{(\zeta-1)!\,\Gamma(\kappa)} \left(\frac{\tilde{m}}{\kappa+\tilde{m}}\right)^{\zeta-1}.$$

The full conditional distribution of $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{m}$ and κ cannot be solved analytically and therefore we use HMC to update these parameters. We have that the partial derivatives are given by:

$$\frac{\partial \log \pi(\tilde{\alpha}, \tilde{\beta}, \tilde{m}, \kappa \mid \cdot)}{\partial \tilde{\alpha}} = \sum_{p=1}^{P} \sum_{t=2}^{T^p} \left[N_{01}^p(t) \times e^{\tilde{\alpha}} \times \frac{\exp\left\{ - e^{\tilde{\alpha}} - e^{\tilde{\beta}} \sum_{c=1}^{n_c^p} x_{t-1}^{[c,\,p]} \right\}}{1 - \exp\left\{ - e^{\tilde{\alpha}} - e^{\tilde{\beta}} \sum_{c=1}^{n_c^p} x_{t-1}^{[c,\,p]} \right\}} \right.$$

$$\left. - N_{00}^p(t) \times e^{\tilde{\alpha}} \right] + \frac{\partial \log \pi_{\alpha}(e^{\tilde{\alpha}})}{\partial \tilde{\alpha}} + 1,$$

$$\frac{\partial \log \pi(\tilde{\alpha}, \tilde{\beta}, \tilde{m}, \kappa \mid \cdot)}{\partial \tilde{\beta}} = \sum_{p=1}^{P} \sum_{t=2}^{T^p} \left[N_{01}^p(t) \times e^{\tilde{\beta}} \times \sum_{c=1}^{n_c^p} x_{t-1}^{[c,\,p]} \times \frac{\exp\left\{ - e^{\tilde{\alpha}} - e^{\tilde{\beta}} \sum_{c=1}^{n_c^p} x_{t-1}^{[c,\,p]} \right\}}{1 - \exp\left\{ - e^{\tilde{\alpha}} - e^{\tilde{\beta}} \sum_{c=1}^{n_c^p} x_{t-1}^{[c,\,p]} \right\}} \right.$$

$$- N_{00}^{p}(t) \times e^{\tilde{\beta}} \times \sum_{c=1}^{n_c^p} x_{t-1}^{[c,\,p]} \Bigg] + \frac{\partial \log \pi_{\beta}(e^{\tilde{\beta}})}{\partial \tilde{\beta}} + 1,$$

$$\frac{\partial \log \pi(\tilde{\alpha}, \tilde{\beta}, \tilde{m}, \kappa \mid \cdot)}{\partial \tilde{m}} = \sum_{p=1}^{P} \sum_{c=1}^{n_c^p} \Bigg[X_1^{[c,\,p]} \times \frac{- \sum_{\zeta=1}^{\zeta_1^{*\,[c,\,p]}-1} \frac{\partial \mathbb{P}(Z = \zeta)}{\partial \tilde{m}}}{1 - \sum_{\zeta=1}^{\zeta_1^{*\,[c,\,p]}-1} \mathbb{P}(Z = \zeta)} - \frac{X_1^{[c,\,p]}}{\tilde{m} + 1}$$

$$+ \frac{\left(1 - X_1^{[c,\,p]}\right)}{\mathbb{P}\left(Z = \zeta_1^{*\,[c,\,p]}\right)} \times \frac{\partial \mathbb{P}\left(Z = \zeta_1^{*\,[c,\,p]}\right)}{\partial \tilde{m}} + \frac{- \sum_{\zeta=1}^{\zeta_{\tau^{[c,\,p]}}^{*\,[c,\,p]}-1} \frac{\partial \mathbb{P}(Z = \zeta)}{\partial \tilde{m}}}{1 - \sum_{\zeta=1}^{\zeta_{\tau^{[c,\,p]}}^{*\,[c,\,p]}-1} \mathbb{P}(Z = \zeta)}$$

$$+ \frac{\left(1 - X_{\tau^{[c,\,p]}}^{[c,\,p]}\right)}{\mathbb{P}\left(Z = \zeta_{\tau^{[c,\,p]}}^{*\,[c,\,p]}\right)} \times \frac{\partial \mathbb{P}\left(Z = \zeta_{\tau^{[c,\,p]}}^{*\,[c,\,p]}\right)}{\partial \tilde{m}} + \sum_{t=2}^{\tau^{[c,\,p]}-1} \frac{\frac{\partial \mathbb{P}\left(Z = \zeta_t^{*\,[c,\,p]}\right)}{\partial \tilde{m}}}{\mathbb{P}\left(Z = \zeta_t^{*\,[c,\,p]}\right)} \Bigg]$$

$$+ \frac{\partial \log \pi_{\tilde{m}}(\tilde{m})}{\partial \tilde{m}},$$

$$\frac{\partial \log \pi(\tilde{\alpha}, \tilde{\beta}, \tilde{m}, \kappa \mid \cdot)}{\partial \kappa} = \sum_{p=1}^{P} \sum_{c=1}^{n_c^p} \Bigg[X_1^{[c,\,p]} \times \frac{- \sum_{\zeta=1}^{\zeta_1^{*\,[c,\,p]}-1} \frac{\partial \mathbb{P}(Z = \zeta)}{\partial \kappa}}{1 - \sum_{\zeta=1}^{\zeta_1^{*\,[c,\,p]}-1} \mathbb{P}(Z = \zeta)}$$

$$+ \frac{\left(1 - X_1^{[c,\,p]}\right)}{\mathbb{P}\left(Z = \zeta_1^{*\,[c,\,p]}\right)} \times \frac{\partial \mathbb{P}\left(Z = \zeta_1^{*\,[c,\,p]}\right)}{\partial \kappa} + \frac{- \sum_{\zeta=1}^{\zeta_{\tau^{[c,\,p]}}^{*\,[c,\,p]}-1} \frac{\partial \mathbb{P}(Z = \zeta)}{\partial \kappa}}{1 - \sum_{\zeta=1}^{\zeta_{\tau^{[c,\,p]}}^{*\,[c,\,p]}-1} \mathbb{P}(Z = \zeta)}$$

$$+ \frac{\left(1 - X_{\tau^{[c,\,p]}}^{[c,\,p]}\right)}{\mathbb{P}\left(Z = \zeta_{\tau^{[c,\,p]}}^{*\,[c,\,p]}\right)} \times \frac{\partial \mathbb{P}\left(Z = \zeta_{\tau^{[c,\,p]}}^{*\,[c,\,p]}\right)}{\partial \kappa} + \sum_{t=2}^{\tau^{[c,\,p]}-1} \frac{\frac{\partial \mathbb{P}\left(Z = \zeta_t^{*\,[c,\,p]}\right)}{\partial \kappa}}{\mathbb{P}\left(Z = \zeta_t^{*\,[c,\,p]}\right)} \Bigg]$$

$$+ \frac{\partial \log \pi_\kappa(\kappa)}{\partial \kappa}.$$

We use a fixed number of leapfrog steps $L = 30$ and adopt the stepsize ϵ during burnin to obtain an acceptance rate of roughly 65% as suggested by Neal (2011).

D.2 Investigating heterogeneity in colonisation rates: simulation studies

We conduct a simulation study using data under epidemic scenarios with various levels of heterogeneity in between-pen colonisation rates. Our goals are to validate parameter estimates obtained through our MCMC algorithm, as well as the effectiveness of IS in detecting the true model under these different scenarios.

Each of the epidemics is simulated in a population of 20 pens with the same structure as the original *E. coli* O157:H7 dataset 1, using an SIS model with parameters fixed to known values. We consider four different scenarios 1-4, and label the corresponding simulated datasets as SD1, SD2, SD3 and SD4 respectively. In each case the data arise using model $\mathcal{M}_k$, $k = 1, 2, 3, 4$, as the true model, where the models are described in Section 5.3.1. The true values of the epidemic model parameters in each setting can be found in Table D.1. We choose these values to reflect our expectations in real world datasets.

Table D.1: Setup of our simulation study. Four datasets are generated, SD1-SD4, with different colonisation rates. The true model that was used to simulate each dataset is given in parentheses.

Simulated	**Parameter**							
Dataset	α_s	α_n	β_s	β_n	m	ν	θ_R	θ_F
SD1 ($\mathcal{M}_1$)	0.015	0.005	0.005	0.015	9	0.1	0.8	0.5
SD2 ($\mathcal{M}_2$)	0.010	0.010	0.005	0.015	9	0.1	0.8	0.5
SD3 ($\mathcal{M}_3$)	0.005	0.015	0.010	0.010	9	0.1	0.8	0.5
SD4 ($\mathcal{M}_4$)	0.009	0.009	0.011	0.011	9	0.1	0.8	0.5

D.2.1 Performance in estimating model parameters

For each dataset SD1, SD2, SD3 and SD4, we fit all 4 candidate models. MCMC is run for 15,000 iterations of which we discard the first 5,000 as a burn-in. From the remaining 10,000 we record samples every 5 iterations and so have a total 1,000 posterior draws. We assign independent prior distributions for the model parameters as follows. In all models, we set $m \sim \text{Ga}(0.01, 0.01)$ and $\nu, \theta_R, \theta_F \sim \text{Beta}(1, 1)$. When included in the model, we assume that $\alpha, \beta, \alpha_s, \alpha_n, \beta_s$ and β_n have a Ga(1,1) prior.

For each scenario, the MCMC output is displayed in the form of marginal posterior distributions in Figure D.1. The figure demonstrates that the posterior distribution of the parameters, under the correctly specified model, is located near to its true value, indicating that the algorithm can successfully recover this parameter. However, when we fit the wrong model to the data, there several points of note.

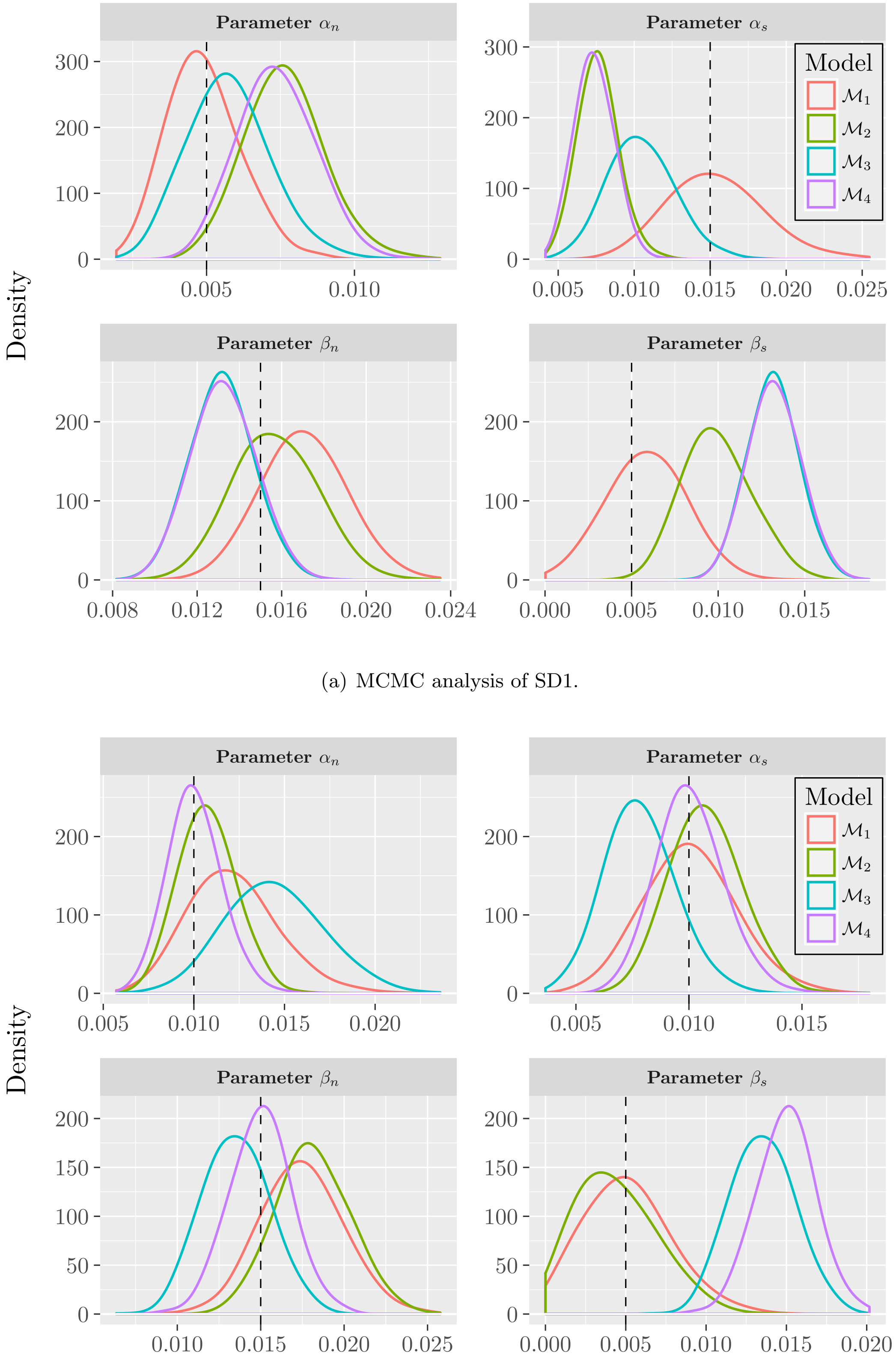
In scenario SD1 (data generated under the full model $\mathcal{M}_1$), both models $\mathcal{M}_2$ and $\mathcal{M}_4$ yield broadly similar densities for the external transmission parameter α. In particular, we see that the posterior distribution for α is located between the true values of α_s and α_n. In a similar manner, models $\mathcal{M}_3$ and $\mathcal{M}_4$ give almost identical densities for the within-pen transmission parameter β, and these estimates are roughly the average of the true values β_s and β_n. Note that models $\mathcal{M}_2 - \mathcal{M}_4$ give misleading inferences regarding the transmission process, when the full model $\mathcal{M}_1$ is the true model. Analogous arguments can be made under scenarios SD2 and SD3. For the former, where we have distinct β_n and β_s values for the North and South pens but a common α, fits under models $\mathcal{M}_3$ and $\mathcal{M}_4$ provide estimates of β in between β_n and β_s leading to inaccurate results. For the latter, we observe that under $\mathcal{M}_2$ and $\mathcal{M}_4$ the estimate of α is within the interval $[\alpha_n, \alpha_s]$. Finally, under scenario SD4 (simple model is the true), all models give similar results for the parameters which are close to the true values.

D.2.2 Performance in determining the true model

In Section D.2.1, we fit all 4 candidate models to datasets SD1-SD4 using a $\text{Ga}(1, 1)$ prior for the colonisation rates. In this section we consider 2 other prior distributions, namely $\text{Ga}(0.1, 0.1)$ and $\text{Ga}(1, 100)$ to check if model comparison produces different results under the 3 alternatives. Hence, for each dataset we fit all distinct combinations of model and prior which results to 12 MCMC outputs per dataset. We then use IS to estimate the posterior probability of each model. This procedure is repeated 40 times using new datasets each time, in order to prevent biases occurring due to the simulated datasets.

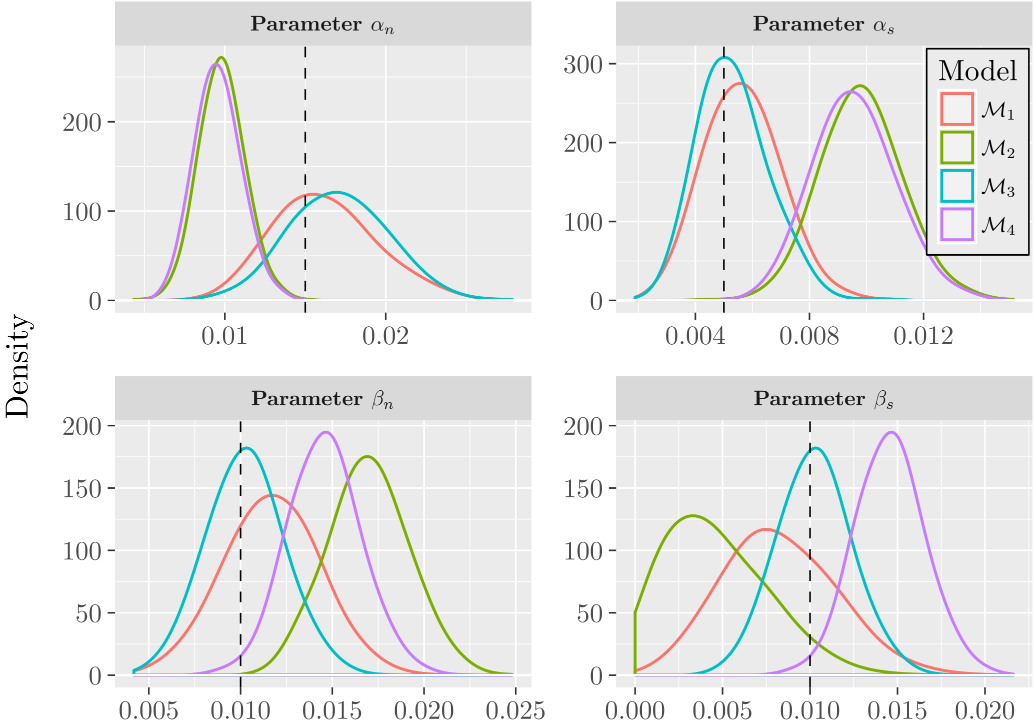
Figure D.2 presents the median posterior model probabilities arranged by the 4 different simulation scenarios. We find that posterior model probabilities are highly sensitive to the choice of prior. More specifically, we see that the $\text{Ga}(1, 1)$ prior tends to favour the simple model $\mathcal{M}_4$ regardless to which one of the 4 models was used to generate the data. The $\text{Ga}(0.1, 0.1)$ show similar behaviour but assigns

Figure D.1: Marginal posterior densities of selected characteristics of simulated data SD1-SD4. The red, green, blue and purple lines correspond to Model $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$, respectively. The dashed lines represent the true parameter values.

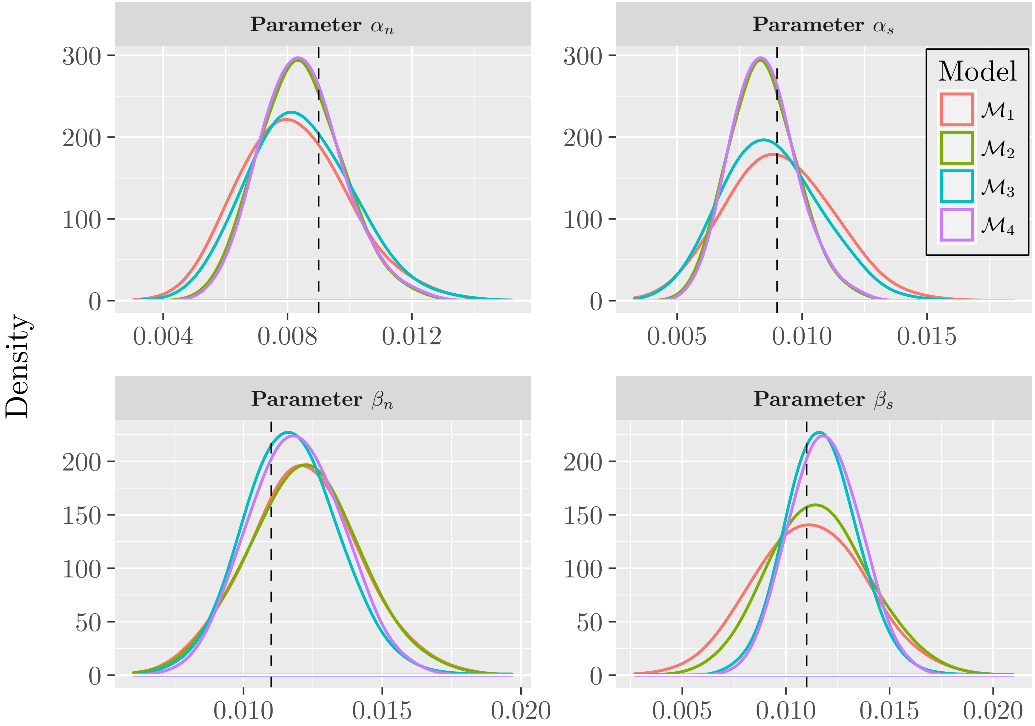


(a) MCMC analysis of SD1.

(b) MCMC analysis of SD2.



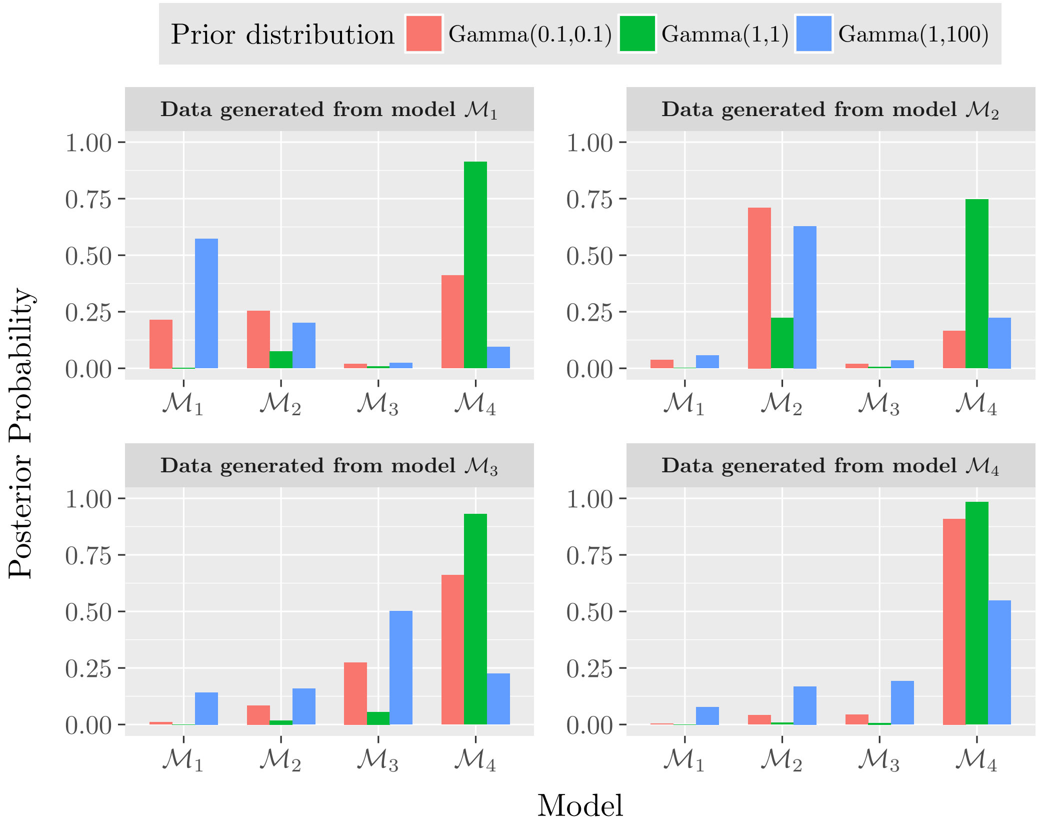
(c) MCMC analysis of SD3.



(d) MCMC analysis of SD4.

higher posterior probabilities to the correct models compared to the Ga$(1, 1)$. The last prior, Ga$(1, 100)$ favours the correct true model under all 4 setups and therefore is used for our real data analysis.

Figure D.2: Posterior probabilities of models $\mathcal{M}_1$-$\mathcal{M}_4$ fit to datasets SD1-SD4. The priors considered for the external and within-pen colonisation rates are the Ga$(0.1, 0.1)$ (red), Ga$(1, 1)$ (green) and Ga$(1, 100)$ (blue).



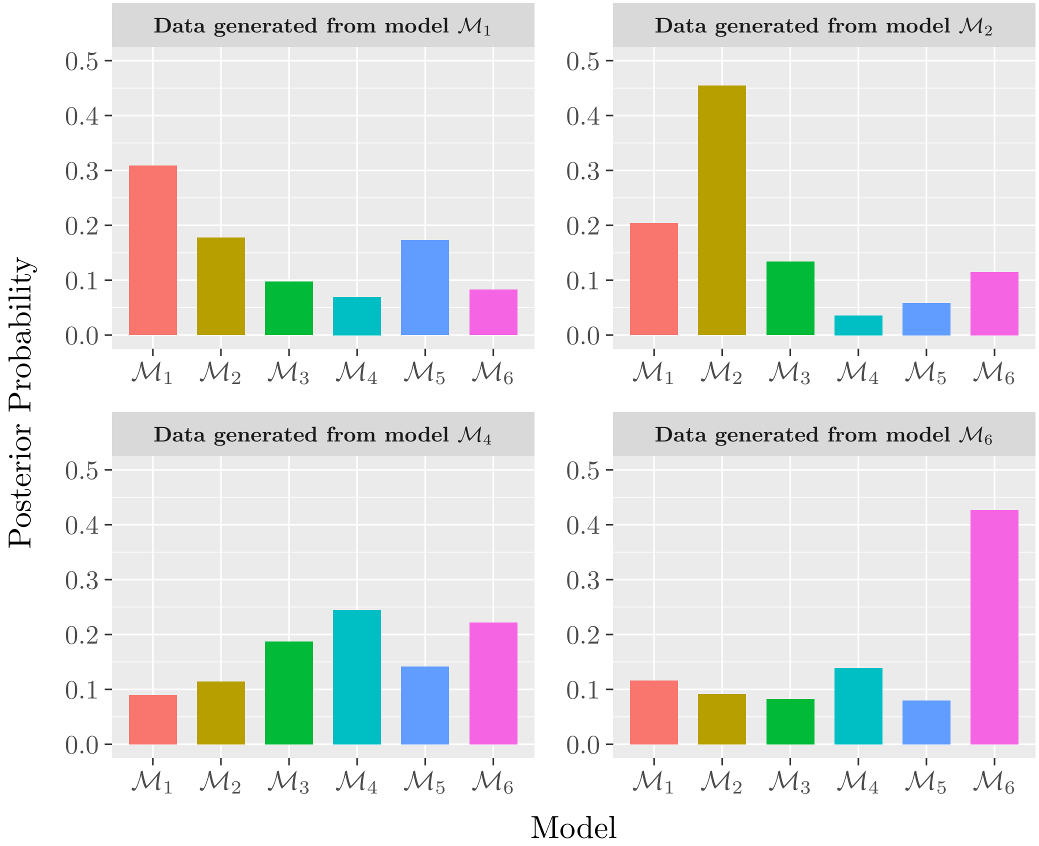
D.3 Investigating transmission between neighbouring pens: simulation studies

D.3.1 Performance in determining the true model

In this section we perform a simulation study to assess the ability of the IS algorithm to determine the true model under scenarios with different forms of interactions between neighbouring pens. The six models that we consider are described in Section 5.4.2.2. The prior for the between pen transmission rate η is a Ga(1, 1000), when this parameter is included in the model. We generate 40 datasets under models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_4$ and $\mathcal{M}_6$, and for each dataset we evaluate the posterior probabilities of the six competing models.

Figure D.3 shows the median posterior probability of each model under the four scenarios. Overall, we observe that our method favours the true model that was used to generate the data. Evidence in favour of the generative model is strong in scenarios 1 ($\mathcal{M}_1$), 2 ($\mathcal{M}_2$) and 4 ($\mathcal{M}_6$), but less strong in scenario 3 ($\mathcal{M}_4$).

FIGURE D.3: Median posterior probabilities of models $\mathcal{M}_1 - \mathcal{M}_6$ using data generated under $\mathcal{M}_1$ (top left), $\mathcal{M}_2$ (top right), $\mathcal{M}_4$ (bottom left) and $\mathcal{M}_6$ (bottom right). In each scenario, the median probability over 40 simulated datasets is presented.



Appendix E

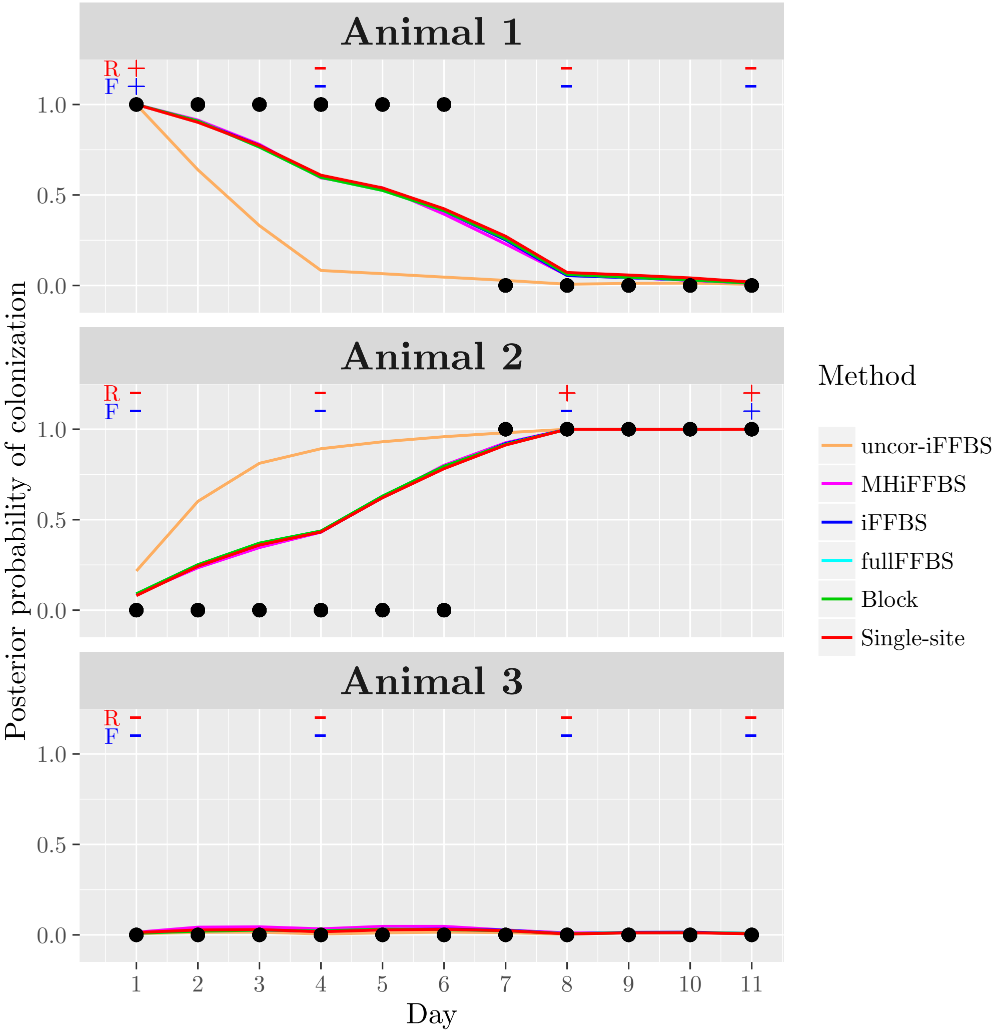
Scalable Inference For Epidemic Models Supplementary Material

E.1 Mixing properties of the uncorrected-iFFBS method with no Metropolis Hastings correction

In this section we provide an example of the poor mixing properties that may occur when we sample directly from the approximated full conditionals of the hidden individual disease states in the CHMM, without correcting with a MH acceptance step. We consider a dataset with a single pen and $C = 3$ individuals for sampling interval of $T = 11$ days and obtain imperfect test results at days $t = 1, 4, 8, 11$. The data are simulated from the Markov model described in Section 3.2 of Chapter 3 where we not allow for transmission of the disease from any other source apart from within-pen transmission. This is achieved by setting the external colonisation rate $\alpha = 0$ which is also the model that we fit. The rest of the model parameters are set according to the posterior median of the real data analysis in Section 3.6.1. We plot the observed data along with the true colonisation states in Figure E.1.

We investigate the efficiency of the methods described in Section 6.4 as well as the standard version of the FFBS (uncorrected-iFFBS) that does not account for approximation of the full conditionals, in the case that update of the hidden states is done animal-by-animal. We evaluate the mixing properties of the different methods by looking at the estimated posterior probability of colonisation for each individual per day over the entire sampling period. Results are also shown in Figure E.1. We see that all methods provide identical results except from the uncorrected-iFFBS method that converges to a different distribution. We further apply the MH correction to the uncorrected-iFFBS method, called MHiFFBS, and see that the method converges to the same values as the rest of the methods, with an acceptance rate of 0.5 for individual 1, 0.75 for individual 2 and 0.95 for individual 3.

Figure E.1: Posterior probability of colonisation for individuals in the simulated dataset of Section E.1, over the entire sampling period of 11 days. Black dots represent the true colonisation states (1 for colonised, 0 for non-carrier). For reference we also show test results taken at days 1, 4, 8, 11.



The poor observed performance can be explained by the example shown in Figure E.2. In this figure we show the sampled hidden disease states at iterations j and $j+1$ of the MCMC using the uncorrected-iFFBS method. In the middle panel, we observe that even though at day 2 the disease has died out, it re-appears at day 3. However, this should be impossible based on the model assumption that does not allow for external transmission of the disease. The reason for this phenomenon is that the sampler ignores the carriage states of other individuals in the next day when it calculates the filtered probabilities. As an example, in the update of animal 1 at day 2 the sampler gives a value of 0; nevertheless, if it accounted for the colonisation state of animal 2 at day 3 then it should give a value of 1 to ensure that there is at least one individual who can transmit the disease to the next day. This is corrected at the update of individual 2 at day 2, which finds that the animal is colonised the following day 3 and hence needs to be colonised at day 2 as well (bottom panel).

Figure E.2: Snapshots of the Gibbs hidden state updates with stand-iFFBS method. Upper panel shows the states after iteration j of the MCMC is complete. Middle panel shows the hidden states after individual 1 has been updated at iteration $j+1$. Finally, the bottom panel represents the same information after the update of the second individual.

$$\begin{array}{cc} & \text{Day} \\ & \begin{array}{ccccccccccc}1&2&3&4&5&6&7&8&9&10&11\end{array} \\ \text{Animal}\ \begin{array}{c}1\\2\\3\end{array} & \begin{bmatrix}1&1&0&0&0&0&0&0&0&0&0\\0&0&1&1&1&1&1&1&1&1&1\\0&0&0&0&0&0&0&0&0&0&0\end{bmatrix}\end{array} \qquad \text{Iteration } j$$

$$\Downarrow$$

$$\begin{array}{cc} & \text{Day} \\ & \begin{array}{ccccccccccc}1&2&3&4&5&6&7&8&9&10&11\end{array} \\ \text{Animal}\ \begin{array}{c}1\\2\\3\end{array} & \begin{bmatrix}1&0&0&0&0&0&0&0&0&0&0\\0&0&1&1&1&1&1&1&1&1&1\\0&0&0&0&0&0&0&0&0&0&0\end{bmatrix}\end{array} \qquad \text{Iteration } j+1$$

Updated animal 1

$$\Downarrow$$

$$\begin{array}{cc} & \text{Day} \\ & \begin{array}{ccccccccccc}1&2&3&4&5&6&7&8&9&10&11\end{array} \\ \text{Animal}\ \begin{array}{c}1\\2\\3\end{array} & \begin{bmatrix}1&0&0&0&0&0&0&0&0&0&0\\0&1&1&1&1&1&1&1&1&1&1\\0&0&0&0&0&0&0&0&0&0&0\end{bmatrix}\end{array} \qquad \text{Iteration } j+1$$

Updated animal 2

Appendix F

Multi-Type Markov Model Supplementary Material

F.1 Gradient expressions for parameter updated with HMC

For notational convenience, we assume that the follow-up period T is the same for all individuals. Let $\tilde{\delta} = \log \delta$, $\tilde{\gamma} = \log \gamma$, $\tilde{\mu}_s = \log \mu_s$, $\tilde{\alpha}_s = \log \alpha_s$, $\tilde{\beta}_s = \log \beta_s$, and $\tilde{\lambda}_s^p(t) = e^{\tilde{\alpha}_s} + \big(\mathbb{1}_{\{p \in \mathcal{S}\}} + e^{\tilde{\gamma}}\, \mathbb{1}_{\{p \in \mathcal{N}\}}\big) e^{\tilde{\beta}_s} I_s^p(t-1)$ for $p = 1, 2, \dots, P$, $s = 1, 2, \dots, n_s$, $t = 2, \dots, T$.

Denote by $N_{rs}^p(t)$ the number of individuals in pen p who were in state r at time $t-1$ and in state s at time t, the corresponding conditional distribution of these parameters given all other parameters takes the form,

$$
\begin{aligned}
&\log \pi(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\mu}}, \tilde{\delta}, \tilde{\gamma} \mid \cdot) = \\
&\quad - \sum_{p=1}^{P} \sum_{t=2}^{T} \left[N_{00}^p(t) \sum_{s=1}^{n_s} \tilde{\lambda}_s^p(t) \right] + \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \left[N_{0j}^p(t) \log \left(\tilde{\lambda}_j^p(t) \right) \right] \\
&\quad - \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \left[N_{0j}^p(t) \log \left(\sum_{s=1}^{n_s} \tilde{\lambda}_s^p(t) \right) \right] \\
&\quad + \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \left[N_{0j}^p(t) \log \left(1 - e^{-\sum_{s=1}^{n_s} \tilde{\lambda}_s^p(t)} \right) \right] \\
&\quad + \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \sum_{\substack{k=1 \\ k \neq j}}^{n_s} \left[N_{jk}^p(t) \log \left(e^{\tilde{\delta}}\, \tilde{\lambda}_k^p(t) \right) \right] \\
&\quad + \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \sum_{\substack{k=0 \\ k \neq j}}^{n_s} \left[N_{jk}^p(t) \log \left(1 - e^{-e^{\tilde{\mu}_j} - \sum_{\substack{s=1 \\ s \neq j}}^{n_s} e^{\tilde{\delta}}\, \tilde{\lambda}_s^p(t)} \right) \right]
\end{aligned}
$$

$$
\begin{aligned}
&- \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \sum_{\substack{k=0 \\ k\neq j}}^{n_s} \left[N_{jk}^{p}(t) \, \log \left(e^{\tilde{\mu}_j} + \sum_{\substack{s=1 \\ s\neq j}}^{n_s} e^{\tilde{\delta}} \, \tilde{\lambda}_s^p(t) \right) \right] \\
&- \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \left[N_{jj}^{p}(t) \left(e^{\tilde{\mu}_j} + \sum_{\substack{s=1 \\ s\neq j}}^{n_s} e^{\tilde{\delta}} \, \tilde{\lambda}_s^p(t) \right) \right] + \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \left[N_{j0}^{p}(t) \, \tilde{\mu}_j \right] \\
&+ \sum_{s=1}^{n_s} \left[\log \pi_\alpha \big(e^{\tilde{\alpha}_s} \big) + \log \pi_\beta \big(e^{\tilde{\beta}_s} \big) + \log \pi_\mu \big(e^{\tilde{\mu}_s} \big) + \tilde{\alpha}_s + \tilde{\beta}_s + \tilde{\mu}_s \right] \\
&+ \log \pi_\delta \big(e^{\tilde{\delta}} \big) + \tilde{\delta} + \log \pi_\gamma \big(e^{\tilde{\gamma}} \big) + \tilde{\gamma}.
\end{aligned}
$$

Then, the partial derivatives are given by

$$
\begin{aligned}
&\frac{\partial \log \pi(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\mu}}, \tilde{\delta}, \tilde{\gamma} \mid \cdot)}{\partial \tilde{\delta}} = \\
&\quad + \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \sum_{\substack{k=1 \\ k\neq j}}^{n_s} \left[N_{jk}^{p}(t) \right] - \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \left[N_{jj}^{p}(t) \sum_{\substack{s=1 \\ s\neq j}}^{n_s} e^{\tilde{\delta}} \, \tilde{\lambda}_s^p(t) \right] \\
&\quad + \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \sum_{\substack{k=0 \\ k\neq j}}^{n_s} \left[N_{jk}^{p}(t) \, \frac{\sum_{\substack{s=1 \\ s\neq j}}^{n_s} \left[e^{\tilde{\delta}} \, \tilde{\lambda}_s^p(t) \right] e^{-e^{\tilde{\mu}_j} - \sum_{\substack{s=1 \\ s\neq j}}^{n_s} e^{\tilde{\delta}} \, \tilde{\lambda}_s^p(t)}}{1 - e^{-e^{\tilde{\mu}_j} - \sum_{\substack{s=1 \\ s\neq j}}^{n_s} e^{\tilde{\delta}} \, \tilde{\lambda}_s^p(t)}} \right] \\
&\quad - \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \sum_{\substack{k=0 \\ k\neq j}}^{n_s} \left[N_{jk}^{p}(t) \, \frac{\sum_{\substack{s=1 \\ s\neq j}}^{n_s} \left[e^{\tilde{\delta}} \, \tilde{\lambda}_s^p(t) \right]}{e^{\tilde{\mu}_j} + \sum_{\substack{s=1 \\ s\neq j}}^{n_s} e^{\tilde{\delta}} \, \tilde{\lambda}_s^p(t) \big]} \right] \\
&\quad + \frac{\partial \log \pi_\delta \big(e^{\tilde{\delta}} \big)}{\partial \tilde{\delta}} + 1
\end{aligned}
\tag{F.1}
$$

and,

$$
\begin{aligned}
&\frac{\partial \log \pi(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\mu}}, \tilde{\delta}, \tilde{\gamma} \mid \cdot)}{\partial \tilde{\gamma}} = \\
&\quad + \sum_{p \in \mathcal{N}} \sum_{t=2}^{T} \sum_{l=1}^{n_s} \left[- N_{00}^{p}(t) \, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\gamma}} + N_{0l}^{p}(t) \, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\gamma}} \, \frac{1}{\tilde{\lambda}_l^p(t)} \right] \\
&\quad - \sum_{p \in \mathcal{N}} \sum_{t=2}^{T} \sum_{l=1}^{n_s} \sum_{j=1}^{n_s} \left[N_{0j}^{p}(t) \, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\gamma}} \, \frac{1}{\sum_{s=1}^{n_s} \tilde{\lambda}_s^p(t)} \right] \\
&\quad + \sum_{p \in \mathcal{N}} \sum_{t=2}^{T} \sum_{l=1}^{n_s} \sum_{j=1}^{n_s} \left[N_{0j}^{p}(t) \, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\gamma}} \, \frac{e^{-\sum_{s=1}^{n_s} \tilde{\lambda}_s^p(t)}}{1 - e^{-\sum_{s=1}^{n_s} \tilde{\lambda}_s^p(t)}} \right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{p\in\mathcal{N}} \sum_{t=2}^{T} \sum_{l=1}^{n_s} \sum_{\substack{j=1\\ j\neq l}}^{n_s} \left[N_{jl}^p(t)\, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\gamma}}\, \frac{1}{\tilde{\lambda}_l^p(t)} \right] \\
&+ \sum_{p\in\mathcal{N}} \sum_{t=2}^{T} \sum_{l=1}^{n_s} \sum_{\substack{j=1\\ j\neq l}}^{n_s} \sum_{\substack{k=0\\ k\neq j}}^{n_s} \left[N_{jk}^p(t)\, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\gamma}}\, \frac{e^{\tilde{\delta}}\, e^{-e^{\tilde{\mu}_j} - \sum_{\substack{s=1\\ s\neq j}}^{n_s} e^{\tilde{\delta}}\, \tilde{\lambda}_s^p(t)}}{1 - e^{-e^{\tilde{\mu}_j} - \sum_{\substack{s=1\\ s\neq j}}^{n_s} e^{\tilde{\delta}}\, \tilde{\lambda}_s^p(t)}} \right] \\
&- \sum_{p\in\mathcal{N}} \sum_{t=2}^{T} \sum_{l=1}^{n_s} \sum_{\substack{j=1\\ j\neq l}}^{n_s} \sum_{\substack{k=0\\ k\neq j}}^{n_s} \left[N_{jk}^p(t)\, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\gamma}}\, \frac{e^{\tilde{\delta}}}{e^{\tilde{\mu}_j} + \sum_{\substack{s=1\\ s\neq j}}^{n_s} e^{\tilde{\delta}}\, \tilde{\lambda}_s^p(t)} \right] \\
&- \sum_{p\in\mathcal{N}} \sum_{t=2}^{T} \sum_{l=1}^{n_s} \sum_{\substack{j=1\\ j\neq l}}^{n_s} \left[N_{jj}^p(t)\, e^{\tilde{\delta}}\, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\gamma}} \right] + \frac{\partial \log \pi_\gamma\big(e^{\tilde{\gamma}}\big)}{\partial \tilde{\gamma}} + 1 \qquad \text{(F.2)}
\end{aligned}
$$

where $\frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\gamma}} = e^{\tilde{\gamma}}\, e^{\tilde{\beta}_l}\, I_l^p(t-1)$. Finally, for each $l = 1, 2, \dots n_s$, we get

$$
\begin{aligned}
&\frac{\partial \log \pi(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\mu}}, \tilde{\delta}, \tilde{\gamma} \mid \cdot)}{\partial \tilde{\mu}_l} = \\
&\quad + \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{\substack{k=0\\ k\neq l}}^{n_s} \left[N_{lk}^p(t)\, \frac{e^{\tilde{\mu}_l}\, e^{-e^{\tilde{\mu}_l} - \sum_{\substack{s=1\\ s\neq l}}^{n_s} e^{\tilde{\delta}}\, \tilde{\lambda}_s^p(t)}}{1 - e^{-e^{\tilde{\mu}_l} - \sum_{\substack{s=1\\ s\neq l}}^{n_s} e^{\tilde{\delta}}\, \tilde{\lambda}_s^p(t)}} \right] \\
&\quad - \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{\substack{k=0\\ k\neq l}}^{n_s} \left[N_{lk}^p(t)\, \frac{e^{\tilde{\mu}_l}}{e^{\tilde{\mu}_l} + \sum_{\substack{s=1\\ s\neq l}}^{n_s} e^{\tilde{\delta}}\, \tilde{\lambda}_s^p(t)} \right] \\
&\quad + \sum_{p=1}^{P} \sum_{t=2}^{T} \left[-\, N_{ll}^p(t)\, e^{\tilde{\mu}_l} + N_{l0}^p(t) \right] + \frac{\partial \log \pi_\mu\big(e^{\tilde{\mu}_l}\big)}{\partial \tilde{\mu}_l} + 1 \qquad \text{(F.3)}
\end{aligned}
$$

and,

$$
\begin{aligned}
&\frac{\partial \log \pi(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\mu}}, \tilde{\delta}, \tilde{\gamma} \mid \cdot)}{\partial \tilde{\alpha}_l} = \\
&\quad + \sum_{p=1}^{P} \sum_{t=2}^{T} \left[-\, N_{00}^p(t)\, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\alpha}_l} + N_{0l}^p(t)\, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\alpha}_l}\, \frac{1}{\tilde{\lambda}_l^p(t)} \right] \\
&\quad - \sum_{p=1}^{P} \sum_{t=2}^{T} \sum_{j=1}^{n_s} \left[N_{0j}^p(t)\, \frac{\partial \tilde{\lambda}_l^p(t)}{\partial \tilde{\alpha}_l}\, \frac{1}{\sum_{s=1}^{n_s} \tilde{\lambda}_s^p(t)} \right]
\end{aligned}
$$

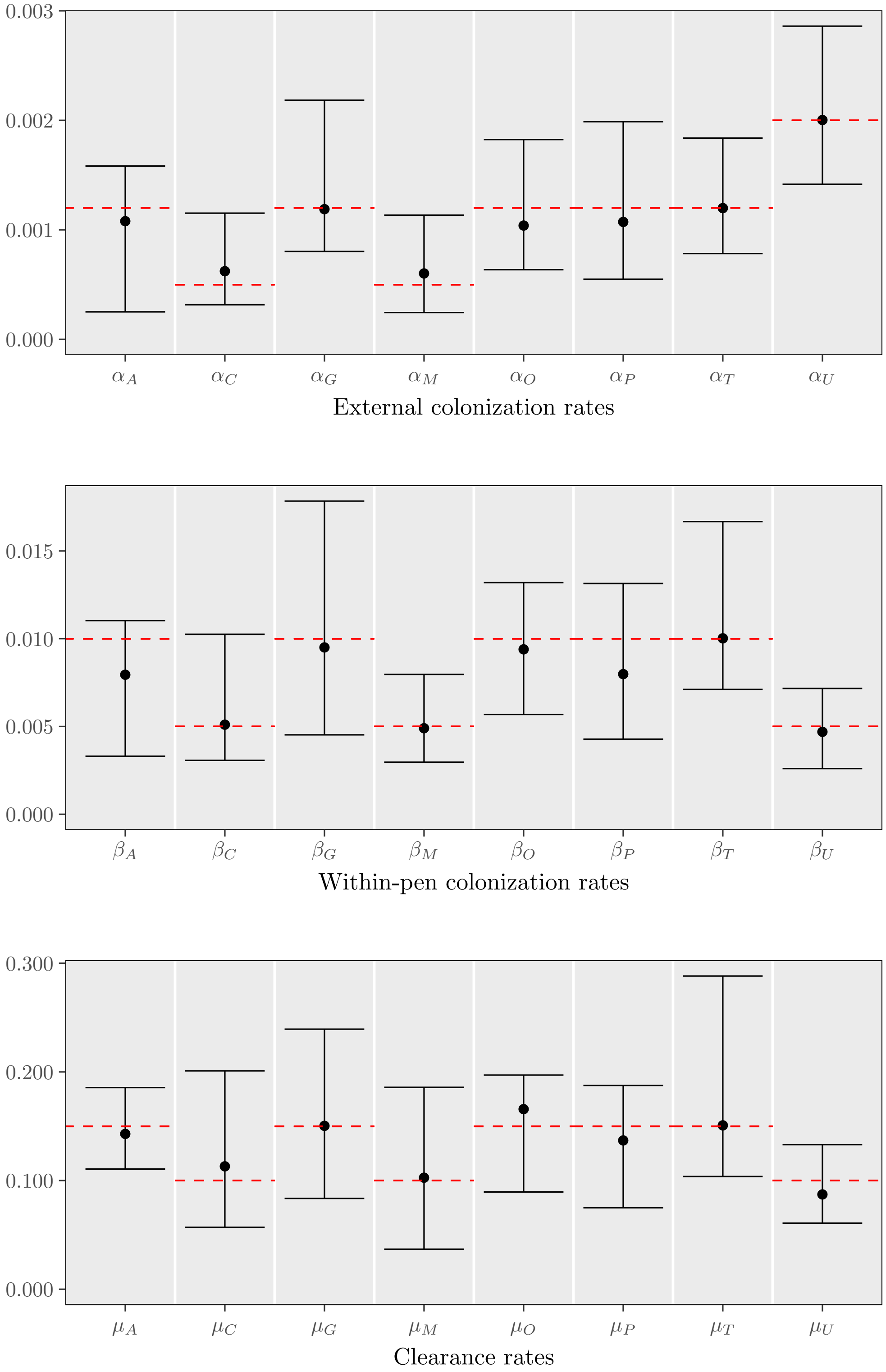
$$
\begin{aligned}
&+\sum_{p=1}^{P}\sum_{t=2}^{T}\sum_{j=1}^{n_s}\left[N_{0j}^{p}(t)\,\frac{\partial\tilde{\lambda}_l^p(t)}{\partial\tilde{\alpha}_l}\,\frac{e^{-\sum_{s=1}^{n_s}\tilde{\lambda}_s^p(t)}}{1-e^{-\sum_{s=1}^{n_s}\tilde{\lambda}_s^p(t)}}\right]\\
&+\sum_{p=1}^{P}\sum_{t=2}^{T}\sum_{\substack{j=1\\ j\neq l}}^{n_s}\left[N_{jl}^{p}(t)\,\frac{\partial\tilde{\lambda}_l^p(t)}{\partial\tilde{\alpha}_l}\,\frac{1}{\tilde{\lambda}_l^p(t)}\right]\\
&+\sum_{p=1}^{P}\sum_{t=2}^{T}\sum_{\substack{j=1\\ j\neq l}}^{n_s}\sum_{\substack{k=0\\ k\neq j}}^{n_s}\left[N_{jk}^{p}(t)\,\frac{\partial\tilde{\lambda}_l^p(t)}{\partial\tilde{\alpha}_l}\,\frac{e^{\tilde{\delta}}\,e^{-e^{\tilde{\mu}_j}-\sum_{\substack{s=1\\ s\neq j}}^{n_s}e^{\tilde{\delta}}\,\tilde{\lambda}_s^p(t)}}{1-e^{-e^{\tilde{\mu}_j}-\sum_{\substack{s=1\\ s\neq j}}^{n_s}e^{\tilde{\delta}}\,\tilde{\lambda}_s^p(t)}}\right]\\
&-\sum_{p=1}^{P}\sum_{t=2}^{T}\sum_{\substack{j=1\\ j\neq l}}^{n_s}\sum_{\substack{k=0\\ k\neq j}}^{n_s}\left[N_{jk}^{p}(t)\,\frac{\partial\tilde{\lambda}_l^p(t)}{\partial\tilde{\alpha}_l}\,\frac{e^{\tilde{\delta}}}{e^{\tilde{\mu}_j}+\sum_{\substack{s=1\\ s\neq j}}^{n_s}e^{\tilde{\delta}}\,\tilde{\lambda}_s^p(t)}\right]\\
&-\sum_{p=1}^{P}\sum_{t=2}^{T}\sum_{\substack{j=1\\ j\neq l}}^{n_s}\left[N_{jj}^{p}(t)\,e^{\tilde{\delta}}\,\frac{\partial\tilde{\lambda}_l^p(t)}{\partial\tilde{\alpha}_l}\right]+\frac{\partial\log\pi_\alpha\big(e^{\tilde{\alpha}_l}\big)}{\partial\tilde{\alpha}_l}+1 \qquad \text{(F.4)}
\end{aligned}
$$

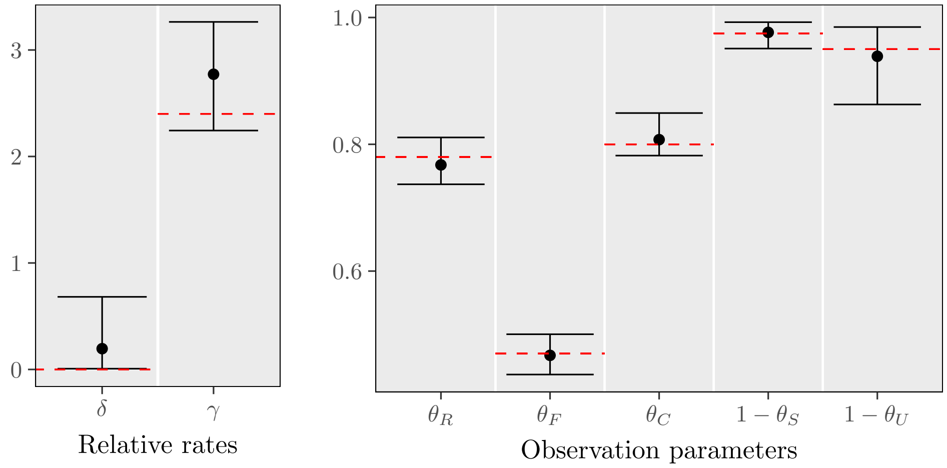
where $\frac{\partial\tilde{\lambda}_l^p(t)}{\partial\tilde{\alpha}_l}=e^{\tilde{\alpha}_l}$. The derivatives with respect to $\tilde{\beta}_l$ are similar; we replace $\frac{\partial\tilde{\lambda}_l^p(t)}{\partial\tilde{\alpha}_l}$ with $\frac{\partial\tilde{\lambda}_l^p(t)}{\partial\tilde{\beta}_l}=\big(\mathbb{1}_{\{p\in\mathcal{S}\}}+e^{\tilde{\gamma}}\,\mathbb{1}_{\{p\in\mathcal{N}\}}\big)e^{\tilde{\beta}_l}\,I_l^p(t-1)$.

F.2 Simulation studies: additional results

In this section we provide additional results for the simulation studies of Section 7.4.2. Marginal posterior summaries over 50 simulated data can be found in Figure F.1. In this scenario we assume that the true data are generated from a model with $\delta=0$ and a full model is estimated.

Figure F.1: Marginal posterior summaries over 50 simulated datasets, each one generated based on the model with $\delta = 0$ and analysed using the full model. Black dots denote the posterior median and error bars indicate the 90% quantile intervals of the 50 posterior medians. Dashed red lines indicate the true value used to simulate the data.





F.3 Real data analysis: additional results

In this section we provide additional results for the analysis of the *E. coli* O157:H7 data of Section 7.5. The estimated serotype-specific prevalence can be found in Figure F.2. Figure F.3 shows posterior traceplots of the model parameters. Figure F.4 presents the results of our prior sensitivity analysis for the rates of colonisation and clearance.

Figure F.2: Posterior serotype-specific prevalence, calculated using the simulated latent carriage process in the *E. coli* dataset 1. The vertical black bars indicate 90% credible intervals.

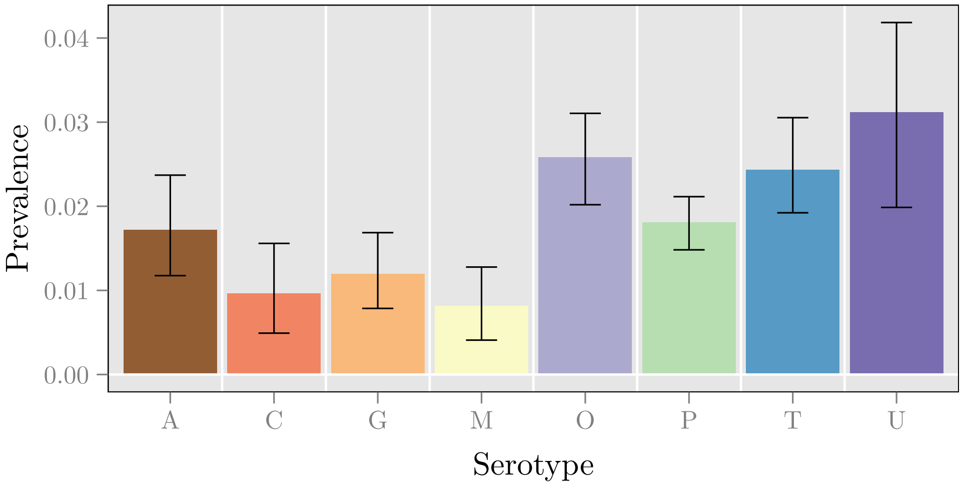
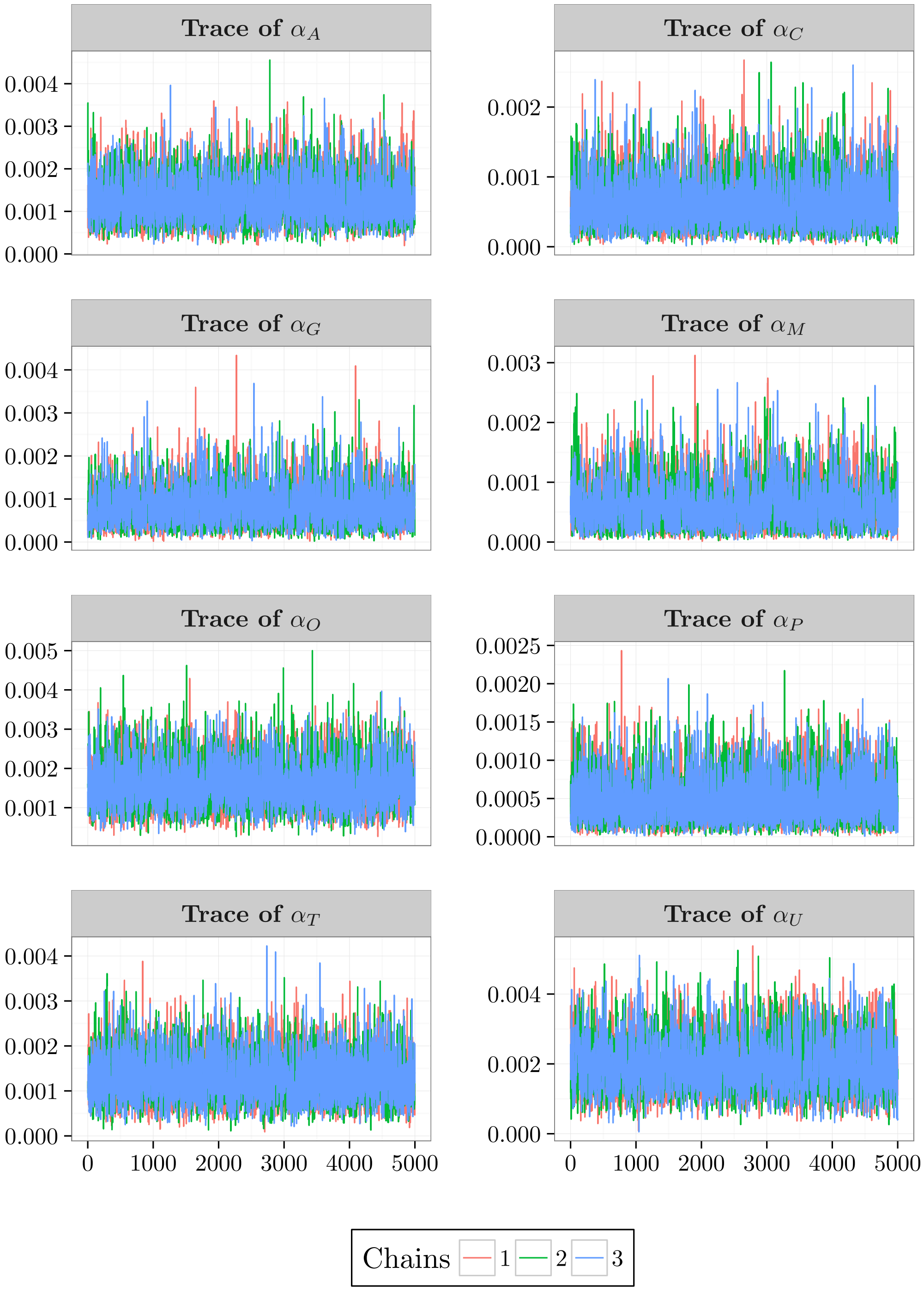
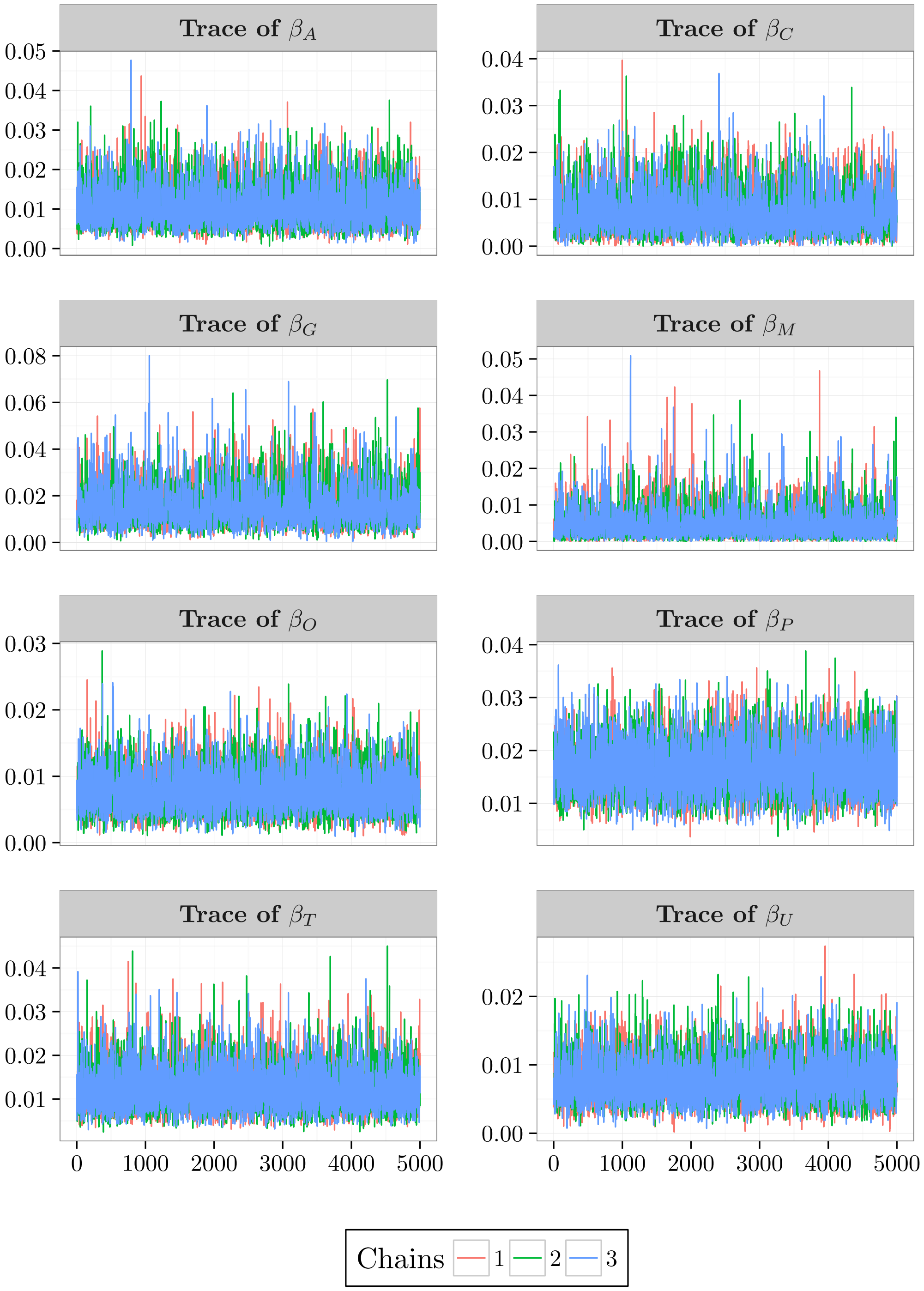


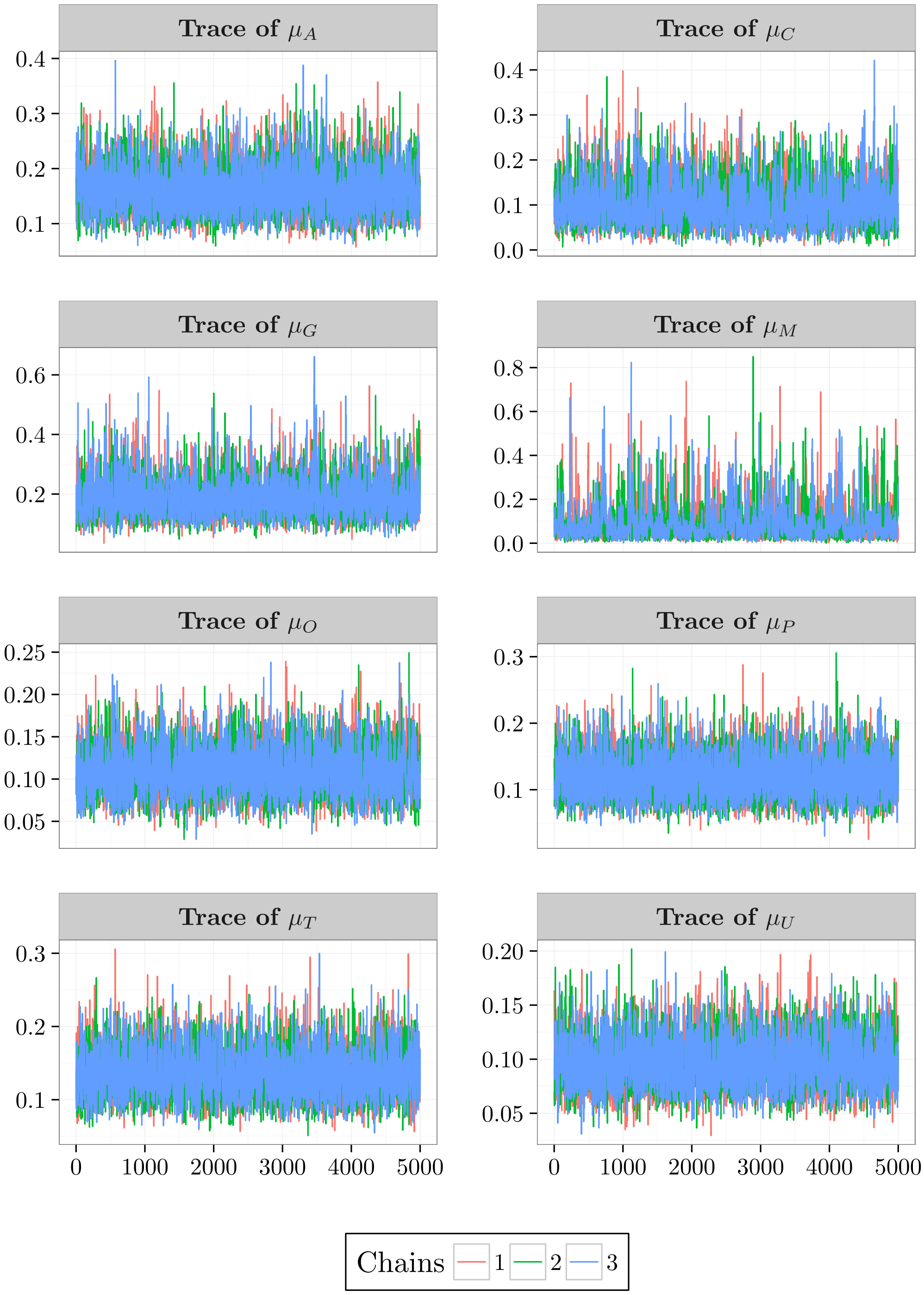
Figure F.3: Posterior traceplots of the multi-serotype model parameters, fit to *E. coli* dataset 1. Results from three independent Markov chains are presented.



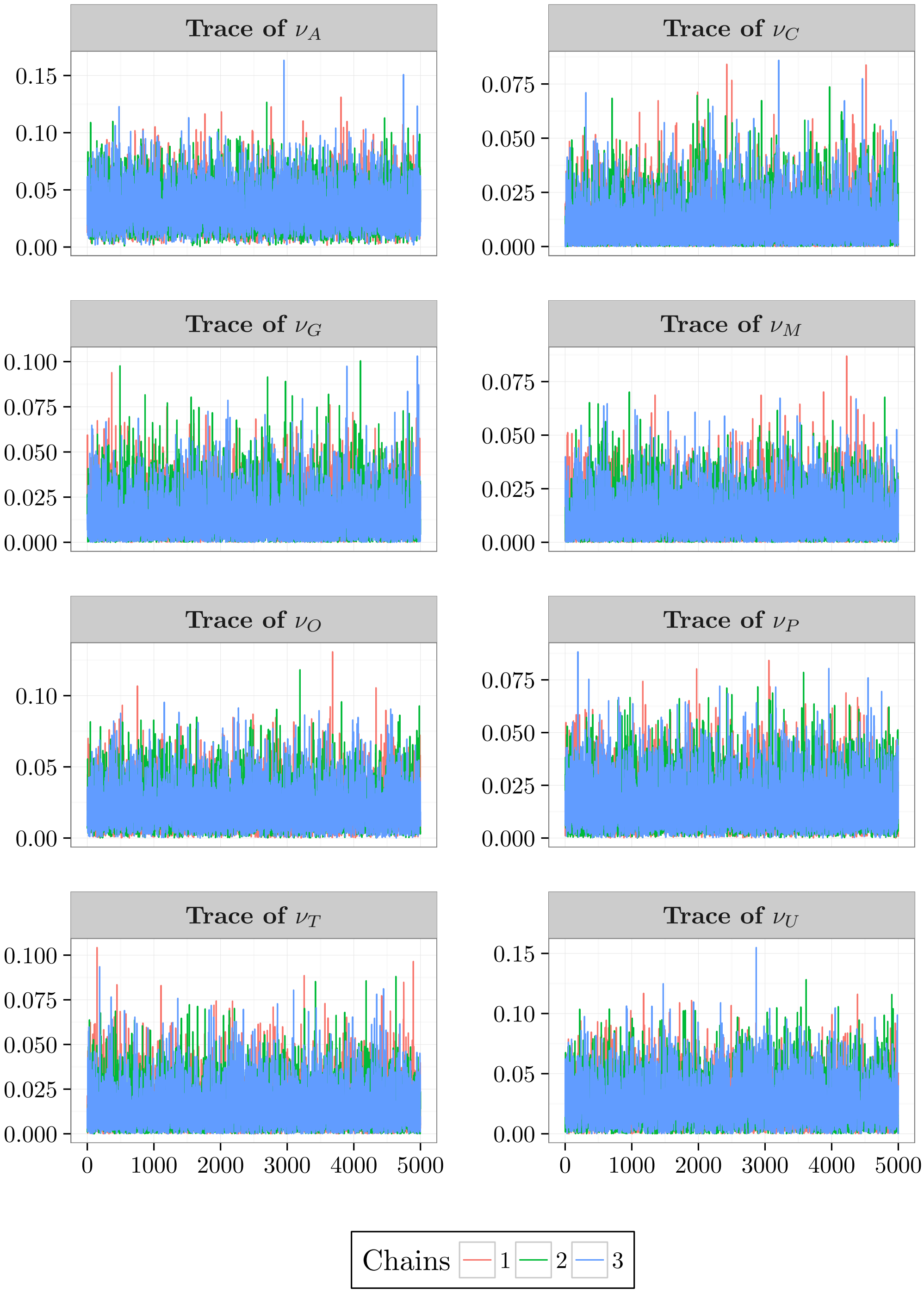
(a) External colonisation rates.



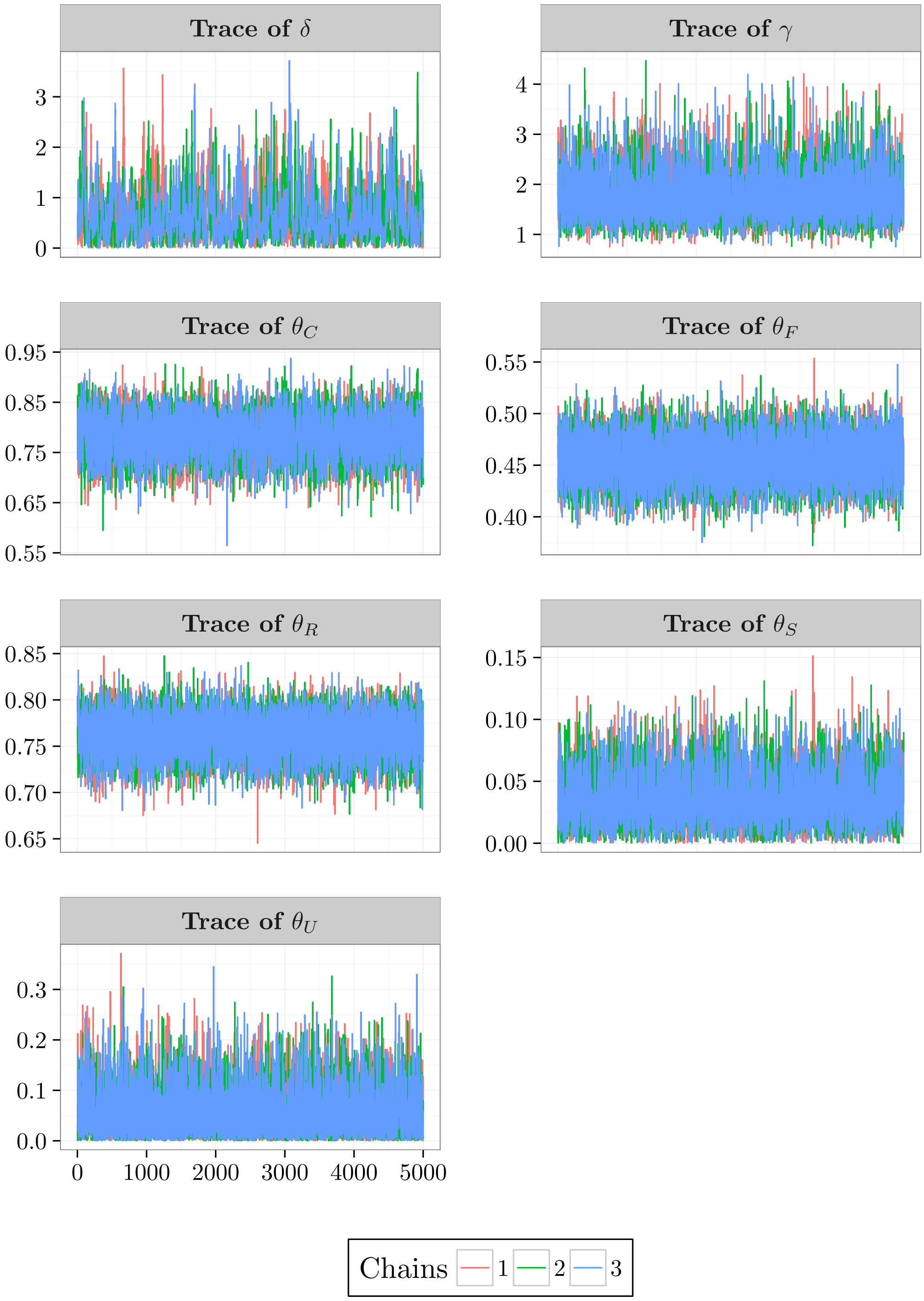
(b) Within-pen colonisation rates.



(c) Clearance rates.

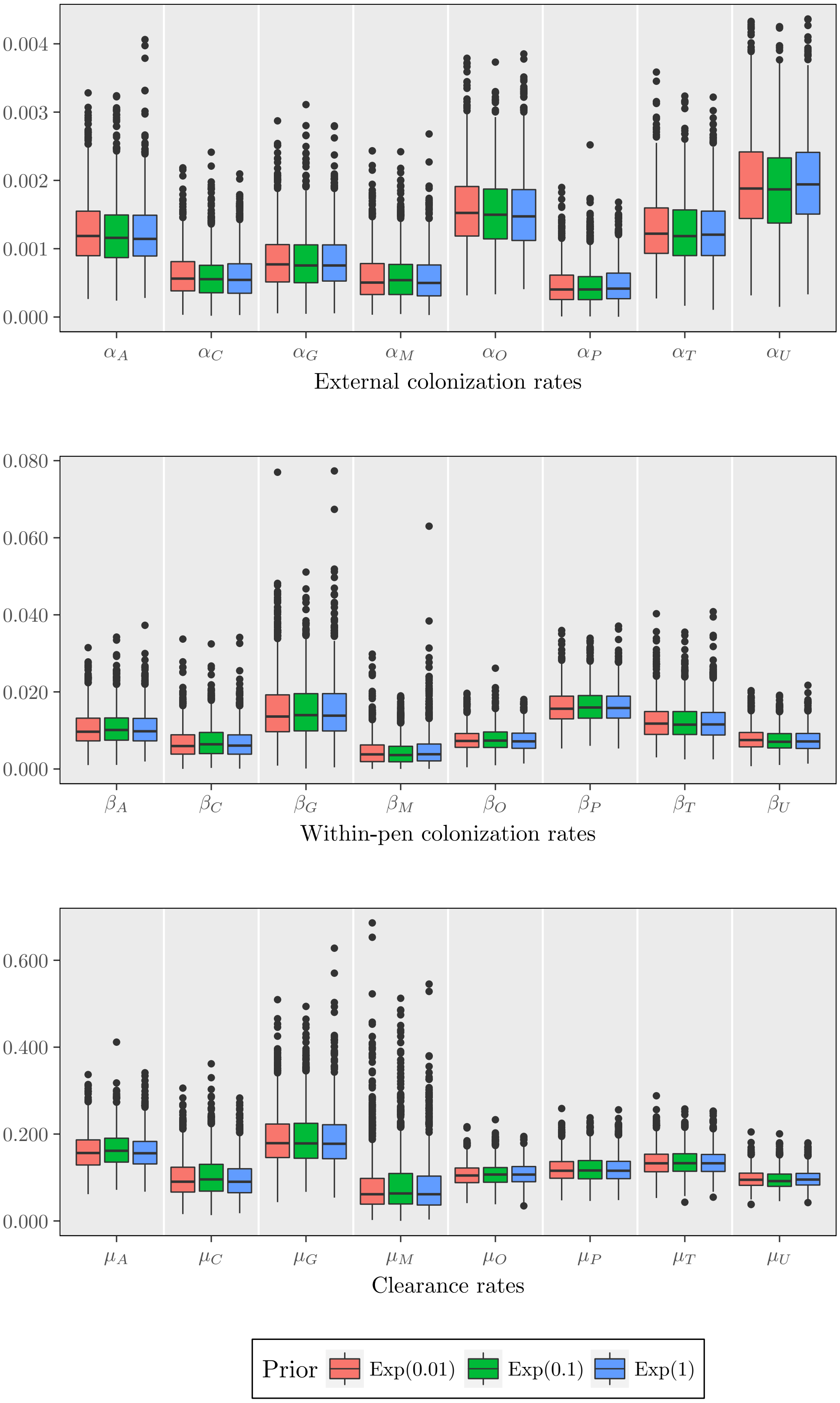


(d) Initial probabilities of carriage.



(e) Relative colonisation rates and observation parameters.

FIGURE F.4: Sensitivity of serotype-specific colonisation and clearance rates to the choice of prior. The prior distributions considered are the Exp(0.01) (red), Exp(0.1) (green) and Exp(1) (blue).



Bibliography

Abbey, H. (1952). An examination of the Reed-Frost theory of epidemics. *Human Biology*, **24**(3), 201–233.

Altman, D. G. (1990). *Practical Statistics for Medical Research*. Chapman and Hall/CRC Texts in Statistical Science.

Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*. Englewood Cliffs, New Jersey: Prentice-Hall.

Anderson, C. J., Wasserman, S., and Crouch, B. (1999). A p^* primer: Logit models for social networks. *Social Networks*, **21**(1), 37–66.

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford Science Publications. Oxford University Press.

Andersson, H. and Britton, T. (1998). Heterogeneity in epidemic models and its effect on the spread of infection. *Journal of Applied Probability*, **35**(3), 651–661.

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*. Lecture Notes in Statistics. Springer New York.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, **37**(2), 697–725.

Auranen, K., Arjas, E., Leino, T., and Takala, A. K. (2000). Transmission of pneumococcal carriage in families: A latent Markov process model for binary longitudinal data. *Journal of the American Statistical Association*, **95**(452), 1044–1053.

Ayscue, P., Lanzas, C., Ivanek, R., and Gröhn, Y. T. (2009). Modeling on-farm *Escherichia coli* O157:H7 population dynamics. *Foodborne Pathogens and Disease*, **6**(4), 461–470.

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and Its Applications*. Mathematics in Medicine Series. Griffin.

Ball, F. and Clancy, D. (1995). The final outcome of an epidemic model with several different types of infective in a large population. *Journal of Applied Probability*, **32**(3), 579–590.

Ball, F., Mollison, D., and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *The Annals of Applied Probability*, **7**(1), 46–89.

Barbour, A. D. (1975). The duration of the closed stochastic epidemic. *Biometrika*, **62**(2), 477–482.

Bartlett, M. S. (1949). Some evolutionary stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, **11**(2), 211–229.

Bartlett, M. S. (1956). Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 81–109. University of California Press, Berkeley.

Bartlett, M. S. (1957). On theoretical models for competitive and predatory biological systems. *Biometrika*, **44**(1/2), 27–42.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, **164**(3), 1139–1160.

Becker, N. (1989). *Analysis of Infectious Disease Data*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability. Taylor and Francis.

Becker, N. and Marschner, I. (1990). The effect of heterogeneity on the spread of disease. In *Stochastic Processes in Epidemic Theory*, pages 90–103. Springer Berlin Heidelberg.

Becker, N. G. (1997). Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases. *Statistical Methods in Medical Research*, **6**(1), 24–37.

Becker, N. G. and Britton, T. (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(2), 287–307.

Becker, N. G. and Hasofer, A. M. (1997). Estimation in epidemics with incomplete observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**(2), 415–429.

Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. *Mémoires de Mathématiques et de Physique, Académie Royale des Sciences, Paris*, pages 1–45.

Brand, M. (1997). Coupled hidden Markov models for modeling interacting processes. Technical report.

Brand, M., Oliver, N., and Pentland, A. (1997). Coupled hidden Markov models for complex action recognition. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 994–999.

Britton, T. (2010). Stochastic epidemic models: A survey. *Mathematical Biosciences*, **225**(1), 24–35.

Britton, T., Kypraios, T., and O'Neill, P. D. (2011). Inference for epidemics with three levels of mixing: Methodology and application to a measles outbreak. *Scandinavian Journal of Statistics*, **38**(3), 578–599.

Brooks, S., Gelman, A., Jones, G., and Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC. CRC Press.

Bureau, A., Shiboski, S., and Hughes, J. P. (2003). Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine*, **22**(3), 441–462.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, **81**(3), 541–553.

Cauchemez, S., Carrat, F., Viboud, C., Valleron, A., and Boëlle, P.-Y. (2004). A Bayesian MCMC approach to study transmission of influenza: Application to household longitudinal data. *Statistics in Medicine*, **23**(22), 3469–3487.

Cauchemez, S., Temime, L., Valleron, A.-J., Varon, E., Thomas, G., Guillemot, D., and Boëlle, P.-Y. (2006). *S. pneumoniae* transmission according to inclusion in conjugate vaccines: Bayesian analysis of a longitudinal follow-up in schools. *BMC Infectious Diseases*, **6**(1), 1–10.

Cauchemez, S., Valleron, A.-J., Boëlle, P.-Y., Flahault, A., and Ferguson, N. M. (2008). Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, **452**(7188), 750–754.

Cernicchiaro, N., Pearl, D. L., McEwen, S. A., Zerby, H. N., Fluharty, F. L., Loerch, S. C., Kauffman, M. D., Bard, J. L., and LeJeune, J. T. (2010). A randomized controlled trial to assess the impact of dietary energy sources, feed supplements, and the presence of super-shedders on the detection of *Escherichia coli* O157:H7 in feedlot cattle using different diagnostic procedures. *Foodborne Pathogens and Disease*, **7**(9), 1071–1081.

Chen, M.-H. (2005). Computing marginal likelihoods from a single MCMC output. *Statistica Neerlandica*, **59**(1), 16–29.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**(432), 1313–1321.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, **75**(1), 79–97.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, **96**(453), 270–281.

Choi, H., Fermin, D., Nesvizhskii, A. I., Ghosh, D., and Qin, Z. S. (2013). Sparsely correlated hidden Markov models with application to genome-wide location studies. *Bioinformatics*, **29**(5), 533–541.

Clancy, D. and O'Neill, P. D. (2007). Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households. *Scandinavian Journal of Statistics*, **34**(2), 259–274.

Clyde, M., Berger, J., Bullard, F., Ford, E., Jefferys, W., Luo, R., Paulo, R., and Loredo, T. (2007). Current challenges in Bayesian model choice. In G. F. Babu and E. D. Feigelson, editors, *Statistical Challenges in Modern Astronomy IV*, volume 371, pages 224–240.

Cobbold, R. and Desmarchelier, P. (2002). Horizontal transmission of Shiga toxin-producing *Escherichia coli* within groups of dairy calves. *Applied and Environmental Microbiology*, **68**(8), 4148–4152.

Cobbold, R. N., Hancock, D. D., Rice, D. H., Berg, J., Stilborn, R., Hovde, C. J., and Besser, T. E. (2007). Rectoanal junction colonization of feedlot cattle by *Escherichia coli* O157:H7 and its association with supershedders and excretion dynamics. *Applied and Environmental Microbiology*, **73**(5), 1563–1568.

Cooper, B. and Lipsitch, M. (2004). The analysis of hospital infection data using hidden Markov models. *Biostatistics*, **5**(2), 223–237.

Cox, D., Hand, D., and Herzberg, A. (2005). *Selected Statistical Papers of Sir David Cox: Volume 1, Design of Investigations, Statistical Methods and Applications*. Analytical Methods for Social Research. Cambridge University Press.

Danon, L., Ford, A. P., House, T., Jewell, C. P., Keeling, M. J., Roberts, G. O., Ross, J. V., and Vernon, M. C. (2011). Networks and the epidemiology of infectious disease. *Interdisciplinary perspectives on Infectious Diseases*, **2011**(Article ID 284909), 1–28.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Dellaportas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, **86**(3), 615–633.

Demiris, N. and O'Neill, P. D. (2005). Bayesian inference for epidemics with two levels of mixing. *Scandinavian Journal of Statistics*, **32**(2), 265–280.

Dong, W., Pentland, A., and Heller, K. A. (2012). Graph-coupled HMMs for modeling the spread of infection. *arXiv preprint arXiv:1210.4864*.

Doucet, A. and Johansen, A. M. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovsky, editors, *The Oxford Handbook of Nonlinear Filtering*, pages 656–704. Oxford University Press.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, **195**(2), 216–222.

Dukic, V., Lopes, H. F., and Polson, N. G. (2012). Tracking epidemics with Google flu trends data and a state-space SEIR model. *Journal of the American Statistical Association*, **107**(500), 1410–1426.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

Erästö, P., Hoti, F., and Auranen, K. (2012). Modeling transmission of multitype infectious agents: Application to carriage of *Streptococcus pneumoniae*. *Statistics in Medicine*, **31**(14), 1450–1463.

Faith, N. G., Shere, J. A., Brosch, R., Arnold, K. W., Ansay, S. E., Lee, M. S., Luchansky, J. B., and Kaspar, C. W. (1996). Prevalence and clonal nature of

Escherichia coli O157:H7 on dairy farms in Wisconsin. *Applied and Environmental Microbiology*, **62**(5), 1519–1525.

Fearnhead, P. and Meligkotsidou, L. (2004). Exact filtering for partially observed continuous time models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(3), 771–789.

Ferens, W. A. and Hovde, C. J. (2011). *Escherichia coli* O157:H7: Animal reservoir and sources of human infection. *Foodborne Pathogens and Disease*, **8**(4), 465–487.

Ferguson, N. M., Galvani, A. P., and Bush, R. M. (2003). Ecological and immunological determinants of influenza evolution. *Nature*, **422**(6930), 428–433.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(3), 589–607.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer New York.

Gansheroff, L. J. and O'Brien, A. D. (2000). *Escherichia coli* O157:H7 in beef cattle presented for slaughter in the U.S.: Higher prevalence rates than previously estimated. *Proceedings of the National Academy of Sciences*, **97**(7), 2959–2961.

Gelfand, A. E. and Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, **56**(3), 501–514.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**(4), 457–472.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**(4), 733–760.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(6), 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, and J. F. M. Smith, editors, *Bayesian Statistics, 4*, pages 169–193. Oxford University Press.

Gibson, G. J. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*, **15**(1), 19–40.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC. Taylor & Francis.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **73**(2), 123–214.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**(4), 711–732.

Greenquist, M. A., Drouillard, J. S., Sargeant, J. M., Depenbusch, B. E., Shi, X., Lechtenberg, K. F., and Nagaraja, T. G. (2005). Comparison of rectoanal mucosal swab cultures and fecal cultures for determining prevalence of *Escherichia coli* O157:H7 in feedlot cattle. *Applied and Environmental Microbiology*, **71**(10), 6431–6433.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, **7**(2), 223–242.

Hamer, W. (1906). Epidemic disease in england. *The Lancet*, (1), 733–739.

Hastie, D. I. and Green, P. J. (2012). Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*, **66**(3), 309–338.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**(1), 97–109.

Hayakawa, Y., O'Neill, P. D., Upton, D., and Yip, P. S. F. (2003). Bayesian inference for a stochastic epidemic model with uncertain numbers of susceptibles of several types. *Australian and New Zealand Journal of Statistics*, **45**(4), 491–502.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, **37**(2), 185–194.

Hoffman, M. and Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**(1), 1593–1623.

Hussain, M., Melegaro, A., Pebody, R. G., George, R., Edmunds, W. J., Talukdar, R., Martin, S. A., Efstratiou, A., and Miller, E. (2005). A longitudinal household study of *Streptococcus pneumoniae* nasopharyngeal carriage in a UK setting. *Epidemiology and Infection*, **133**(5), 891–898.

Hussein, H. and Sakuma, T. (2005). Prevalence of Shiga toxin-producing *Escherichia coli* in dairy cattle and their products. *Journal of Dairy Science*, **88**(2), 450–465.

Ionides, E. L., Bretó, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, **103**(49), 18438–18443.

Isham, V. (2005). Stochastic models for epidemics. In A. Davison, Y. Dodge, and N. Wermuth, editors, *Celebrating Statistics: Papers in Honour of Sir David Cox on His 80th Birthday*, chapter 1, pages 27–54. Oxford University Press.

Jewell, C. P., Kypraios, T., Neal, P., and Roberts, G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, **4**(3), 465–496.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(2), 143–170.

Karagiannis, G. and Andrieu, C. (2013). Annealed importance sampling reversible jump MCMC algorithms. *Journal of Computational and Graphical Statistics*, **22**(3), 623–648.

Karch, H., Tarr, P. I., and Bielaszewska, M. (2005). Enterohaemorrhagic *Escherichia coli* in human medicine. *International Journal of Medical Microbiology*, **295**(6-7), 405–418.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **115**(772), 700–721.

Knock, E. S. and O'Neill, P. D. (2014). Bayesian model choice for epidemic models with two levels of mixing. *Biostatistics*, **15**(1), 46–59.

Kwon, J. and Murphy, K. (2000). Modeling freeway traffic with coupled HMMs. Technical report, University of California, Berkeley.

Kypraios, T., O'Neill, P. D., Huang, S. S., Rifas-Shiman, S. L., and Cooper, B. S. (2010). Assessing the role of undetected colonization and isolation precautions in reducing Methicillin-Resistant *Staphylococcus aureus* transmission in intensive care units. *BMC Infectious Diseases*, **10**(1), 1–10.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., and Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data: Hidden Markov models and extensions. *Ecology*, **93**(11), 2336–2342.

Lawson, A. (2013). *Statistical Methods in Spatial Epidemiology*. Wiley Series in Probability and Statistics. Wiley.

Lawson, A., Banerjee, S., Haining, R., and Ugarte, M. (2016). *Handbook of Spatial Epidemiology*. Chapman & Hall/CRC. CRC Press.

Lee, X. J., Drovandi, C. C., and Pettitt, A. N. (2015). Model choice problems using approximate Bayesian computation with applications to pathogen transmission data sets. *Biometrics*, **71**(1), 198–207.

Lekone, P. E. and Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, **62**(4), 1170–1177.

Liu, W.-C., Jenkins, C., Shaw, D. J., Matthews, L., Pearce, M. C., Low, J. C., Gunn, G. J., Smith, H. R., Frankel, G., and Woolhouse, M. E. J. (2005). Modelling the epidemiology of Verocytotoxin-producing *Escherichia coli* serogroups in young calves. *Epidemiology and Infection*, **133**(3), 449–458.

Longini, J. I. M. and Koopman, J. S. (1982). Household and community transmission parameters from final distributions of infections in households. *Biometrics*, **38**(1), 115–126.

Low, J. C., McKendrick, I. J., McKechnie, C., Fenlon, D., Naylor, S. W., Currie, C., Smith, D. G. E., Allison, L., and Gally, D. L. (2005). Rectal carriage of Enterohemorrhagic *Escherichia coli* O157 in slaughtered cattle. *Applied and Environmental Microbiology*, **71**(1), 93–97.

MacDonald, I. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hall/CRC. Taylor & Francis.

Matthews, L., Low, J. C., Gally, D. L., Pearce, M. C., Mellor, D. J., Heesterbeek, J. A. P., Chase-Topping, M., Naylor, S. W., Shaw, D. J., Reid, S. W. J., Gunn, G. J., and Woolhouse, M. E. J. (2006a). Heterogeneous shedding of *Escherichia coli* O157 in cattle and its implications for control. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(3), 547–552.

Matthews, L., Mckendrick, I. J., Ternent, H., Gunn, G. J., Synge, B., and Woolhouse, M. E. J. (2006b). Super-shedding cattle and the transmission dynamics of *Escherichia coli* O157. *Epidemiology and Infection*, **134**(1), 131–142.

McKendrick, A. G. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, **44**, 98–130.

McKinley, T., Cook, A. R., and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, **5**(1), 24.

McKinley, T. J., Ross, J. V., Deardon, R., and Cook, A. R. (2014). Simulation-based bayesian inference for epidemic models. *Computational Statistics and Data Analysis*, **71**, 434–447.

Mead, P. S., Finelli, L., Lambert-Fair, M. A., Champ, D., Townes, J., Hutwagner, L., Barrett, T., Spitalny, K., and Mintz, E. (1997). Risk factors for sporadic infection with *Escherichia coli* O157:H7. *Archives of Internal Medicine*, **157**(2), 204–208.

Melegaro, A., Gay, N., and Medley, G. F. (2004). Estimating the transmission parameters of pneumococcal carriage in households. *Epidemiology and Infection*, **132**(3), 433–441.

Melegaro, A., Choi, Y., Pebody, R., and Gay, N. (2007). Pneumococcal carriage in United Kingdom families: Estimating serotype-specific transmission parameters from longitudinal data. *American Journal of Epidemiology*, **166**(2), 228–235.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, **6**(4), 831–860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092.

Morton, A. and Finkenstädt, B. F. (2005). Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(3), 575–594.

Nåsell, I. (2002). Stochastic models of some endemic infections. *Mathematical Biosciences*, **179**(1), 1–19.

Natarajan, P. and Nevatia, R. (2007). Coupled hidden semi Markov models for activity recognition. In *Motion and Video Computing, 2007. WMVC '07. IEEE Workshop on*, pages 10–10.

Neal, P. J. and Roberts, G. O. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics*, **5**(2), 249–261.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, and X. Meng, editors, *Handbook of Markov Chain Mote Carlo*, chapter 5, pages 113–162. Chapman & Hall/CRC.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, **56**(1), 3–48.

Numminen, E., Cheng, L., Gyllenberg, M., and Corander, J. (2013). Estimating the transmission dynamics of *Streptococcus pneumoniae* from strain prevalence data. *Biometrics*, **69**(3), 748–757.

O'Neill, P., Balding, D., Becker, N., Eerola, M., and Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **49**(4), 517–542.

O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences*, **180**(1-2), 103–114.

O'Neill, P. D. and Becker, N. G. (2001). Inference for an epidemic when susceptibility varies. *Biostatistics*, **2**(1), 99–108.

O'Neill, P. D. and Kypraios, T. (2014). Bayesian model choice via mixture distributions with application to epidemics and population process models. *arXiv preprint arXiv:1411.7888*.

O'Neill, P. D. and Marks, P. J. (2005). Bayesian model choice and infection route modelling in an outbreak of Norovirus. *Statistics in Medicine*, **24**(13), 2011–2024.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **162**(1), 121–129.

Raftery, A. E. and Lewis, S. (1992). How many iterations in the Gibbs sampler? In A. P. D. J. M. Bernardo, J. O. Berger and A. F. M. Smith, editors, *In Bayesian Statistics 4*, pages 765–776. Oxford University Press.

Rahn, K., Renwick, S., Johnson, R., Wilson, J., Clarke, R., Alves, D., McEwen, S., Lior, H., and Spika, J. (1997). Persistence of *Escherichia coli* O157:H7 in dairy cattle and the dairy farm environment. *Epidemiology and Infection*, **119**(2), 251–259.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **27**(2/4), 311–324.

Rice, D. H., Sheng, H. Q., Wynia, S. A., and Hovde, C. J. (2003). Rectoanal mucosal swab culture is more sensitive than fecal culture and distinguishes *Escherichia coli* O157:H7-colonized cattle and those transiently shedding the same organism. *Journal of Clinical Microbiology*, **41**(11), 4924–4929.

Rida, W. N. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *Journal of the Royal Statistical Society: Series B (Methodological)*, **53**(1), 269–283.

Riley, L. W., Remis, R. S., Helgerson, S. D., McGee, H. B., Wells, J. G., Davis, B. R., Hebert, R. J., Olcott, E. S., Johnson, L. M., Hargrett, N. T., Blake, P. A., and Cohen, M. L. (1983). Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *New England Journal of Medicine*, **308**(12), 681–685.

Ripley, B. D. (1987). *Stochastic Simulation*. Wiley Series in Probability and Statistics. Wiley.

Robert, C. (2001). *The Bayesian Choice*. Springer Texts in Statistics. 2nd edition, Springer-Verlag, New Work.

Roberts, G. and Tweedie, R. (2005). *Understanding Monte Carlo Markov Chain*. Springer Series in Statistics. Springer Verlag.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**(1), 255–268.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, **18**(2), 349–367.

Ross, R. (1911). *The prevention of malaria, 2nd edition*. London: John Murray.

Ross, R. (1916). An application of the theory of probabilities to the study of a priori pathometry. Part I. *Proceedings of the Royal Society of London*, **A92**(638), 204–230.

Ross, R. and Hudson, H. P. (1917a). An application of the theory of probabilities to the study of a priori pathometry. Part II. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 93, pages 212–225. The Royal Society.

Ross, R. and Hudson, H. P. (1917b). An application of the theory of probabilities to the study of a priori pathometry. Part III. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 93, pages 225–240. The Royal Society.

Saul, L. K. and Jordan, M. I. (1999). Mixed memory Markov models: Decomposing complex stochastic processes as mixtures of simpler ones. *Machine Learning*, **37**(1), 75–87.

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**(457), 337–351.

Shere, J., Kaspar, C., Bartlett, K., Linden, S., Norell, B., Francey, S., and Schaefer, D. (2002). Shedding of *Escherichia coli* O157:H7 in dairy cattle housed in a confined environment following waterborne inoculation. *Applied and Environmental Microbiology*, **68**(4), 1947–1954.

Sherlock, C., Xifara, T., Telfer, S., and Begon, M. (2013). A coupled hidden Markov model for disease interactions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**(4), 609–627.

Skilling, J. (2004). Nested sampling. *Bayesian inference and maximum entropy methods in science and engineering*, **735**(1), 395–405.

Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., *et al.* (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459**(7250), 1122–1125.

Smith, T. and Vounatsou, P. (2003). Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models. *Statistics in Medicine*, **22**(10), 1709–1724.

Spencer, S. E. F., Besser, T. E., Cobbold, R. N., and French, N. P. (2015). 'Super' or just 'above average'? Supershedders and the transmission of *Escherichia coli* O157:H7 among feedlot cattle. *Journal of The Royal Society Interface*, **12**(110), 20150446.

Streftaris, G. and Gibson, G. J. (2004). Bayesian inference for stochastic epidemics in closed populations. *Statistical Modelling*, **4**(1), 63–75.

Sun, L., Lee, C., and Hoeting, J. A. (2015). Parameter inference and model selection in deterministic and stochastic dynamical models via approximate Bayesian computation: Modeling a wildlife epidemic. *Environmetrics*, **26**(7), 451–462.

Tenover, F., Arbeit, R., Goering, R., Murray, B., Persing, D., Pfaller, M., and Weinstein, R. (1997). How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: A review for healthcare epidemiologists. *Infection Control and Hospital Epidemiology*, **18**(6), 426–439.

Teunis, P., Takumi, K., and Shinagawa, K. (2004). Dose response for infection by *Escherichia coli* O157:H7 from outbreak data. *Risk Analysis*, **24**(2), 401–407.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**(393), 82–86.

Toft, N., Jørgensen, E., and Højsgaard, S. (2005). Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. *Preventive Veterinary Medicine*, **68**(1), 19–33.

Tokdar, S., Xi, P., Kelly, R. C., and Kass, R. E. (2010). Detection of bursts in extracellular spike trains using hidden semi-Markov point process models. *Journal of Computational Neuroscience*, **29**(1), 203–212.

Toni, T. and Stumpf, M. P. H. (2010). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, **26**(1), 104–110.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, **6**(31), 187–202.

Touloupou, P., Alzahrani, N., Neal, P., Spencer, S. E., and McKinley, T. J. (2015). Model comparison with missing data using MCMC and importance sampling. *arXiv preprint arXiv:1512.04743*.

Turner, J., Begon, M., Bowers, R. G., and French, N. P. (2003). A model appropriate to the transmission of a human food-borne pathogen in a multigroup managed herd. *Preventive Veterinary Medicine*, **57**(4), 175–198.

Turner, J., Bowers, R. G., Begon, M., Robinson, S. E., and French, N. P. (2006). A semi-stochastic model of the transmission of *Escherichia coli* O157 in a typical UK dairy herd: Dynamics, sensitivity analysis and intervention/prevention strategies. *Journal of Theoretical Biology*, **241**(4), 806–822.

Turner, J., Bowers, R. G., Clancy, D., Behnke, M. C., and Christley, R. M. (2008). A network model of *E. coli* O157 transmission within a typical UK dairy herd: The effect of heterogeneity and clustering on the prevalence of infection. *Journal of Theoretical Biology*, **254**(1), 45–54.

UNAIDS (2015). Aids by the numbers 2015. Technical report.

Weiss, G. H. and Dishon, M. (1971). On the asymptotic behavior of the stochastic and deterministic models of an epidemic. *Mathematical Biosciences*, **11**(3-4), 261–265.

Wells, J., Shipman, L., Greene, K., Sowers, E., Green, J., Cameron, D., Downes, F., Martin, M., Griffin, P., and Ostroff, S. (1991). Isolation of *Escherichia coli* serotype O157: H7 and other Shiga-like-toxin-producing *E. coli* from dairy cattle. *Journal of Clinical Microbiology*, **29**(5), 985–989.

Whittle, P. (1955). The outcome of a stochastic epidemic - a note on Bailey's paper. *Biometrika*, **42**(1-2), 116–122.

Worby, C. J., O'Neill, P. D., Kypraios, T., Robotham, J. V., De Angelis, D., Cartwright, E. J., Peacock, S. J., and Cooper, B. S. (2016). Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The Annals of Applied Statistics*, **10**(1), 395.

Zhong, S. and Ghosh, J. (2002). Hmms and coupled hmms for multi-channel eeg classification. In *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on*, volume 2, pages 1154–1159.

Zhou, Y., Johansen, A. M., and Aston, J. A. (2016). Towards automatic model comparison: An adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, **25**(3), 701–726.